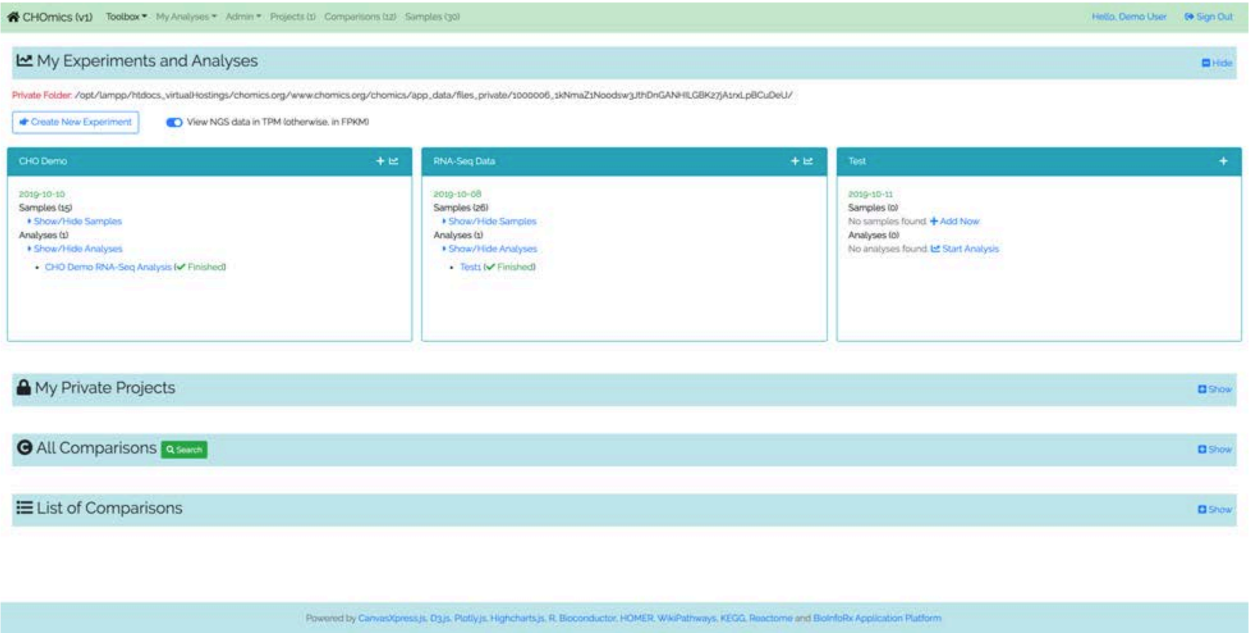
CHOmics Tutorial

Version 1.0, Feb, 2020



Private Folder: /opt/lampp/htdocs_virtualHostings/chomics.org/www.chomics.org/chomics/app_data/files_private/1000006_1kNmaZ1Noodsw3JthDnGANHILGBK27jA1rxLpBCuDeU/

View NGS data in TPM (otherwise, in FPKM)

CHO Demo: 2019-10-10, Samples (15), Analyses (1), CHO Demo RNA-Seq Analysis (✓ Finished)

RNA-Seq Data: 2019-10-08, Samples (26), Analyses (1), Test1 (✓ Finished)

Test: 2019-10-11, Samples (0), No samples found. Analyses (0), No analyses found.

Powered by CanvasXpress.js, D3.js, Plotly.js, Highcharts.js, R, Bioconductor, HOMER, WikiPathways, KEGG, Reactome and BioInfoRx Application Platform

http://chomics.org

user:demo@bioinforx.com

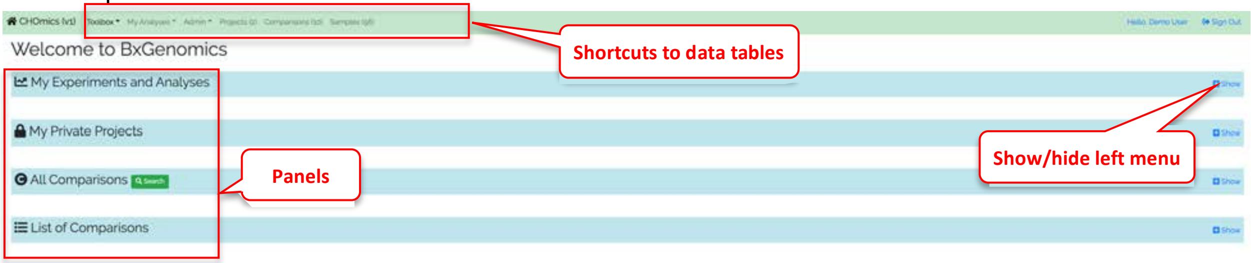
password:CHO_demo

From the login page, you can use your email to register an account that is recommended, as you will be able to save results and upload your own data. Otherwise just use guest account to view public data.

Contents

1. Overview of CHOmics

There are several panels stacked in the main interface. The recent experiment and projects are listed in the panel separately for quick access. You can also access them and other functions from the shortcuts at the top menu bar.



1.1 Menu Bar

In top menu bar, several shortcuts are listed for quick access of functions including: Toolbox, My Analysis, and Admin, Projects, Comparisons and Samples.

‘Toolbox’ contains a list of functional modules including: ‘Import Project Data’,’Gene Expression Analysis’, ’Comparison-based Analysis’, ’Pathway Visualization’ and ‘Other tools’.

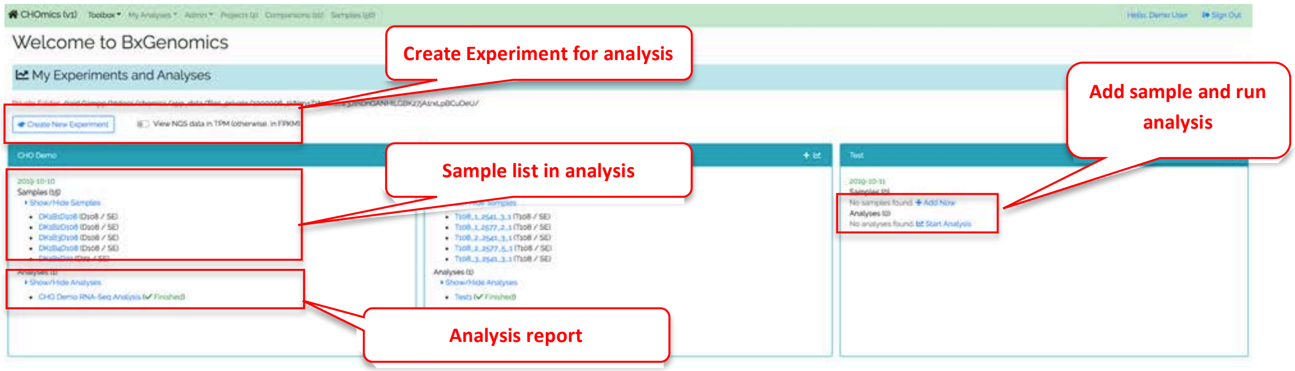
‘My Analysis’ provides quick access to the information of all ‘Experiments’, ‘Samples’ and ‘Analysis’.

‘Admin’ allows the users to manage the data files from private folder, shared folder and overview the platforms applied to all data sets.

‘Projects’, ‘Comparisons’ and ‘Samples’ all provide searching function and access to specific project, comparison and sample respectively.

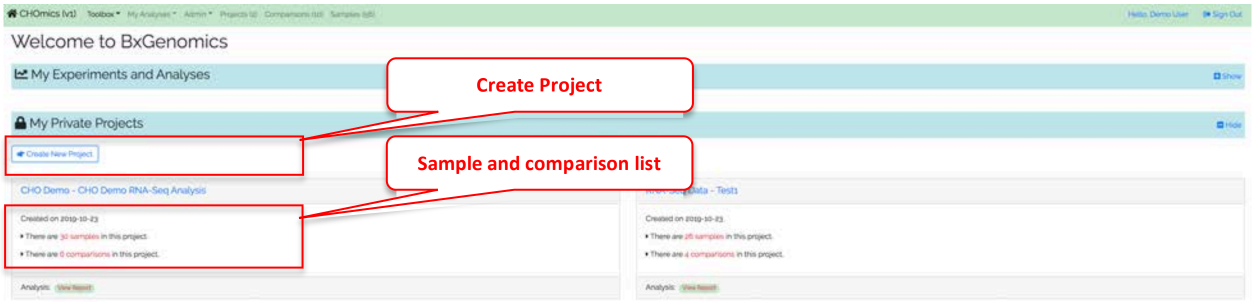
1.2 Experiments and Analyses

Experiment is designed for running the built-in RNA sequencing pipeline on the raw sequencing data. Once the experiment is created, users can upload raw fastq files and sample meta information, and then launch the built-in pipeline for analysis. After the analysis is completed, the analysis report is generated and the results can be exported as one ‘Project’ for visualization and cross-project comparison.



1.3 Projects

The project is used to perform data mining and data visualization. Users can either import analysis report from ‘Experiment’ or upload pre-processed data to create a project. In the project, users can easily explore different features of the data (e.g, Gene expression profiling, sample clustering, PCA, differential expression genes and pathways, etc), compare the analysis with the other projects or perform the meta-analysis by combining multiple projects.



Welcome to BxGenomics

My Experiments and Analyses

My Private Projects

Create Project

Sample and comparison list

CHO Demo - CHO Demo RNA-Seq Analysis

Created on 2019-10-23
- There are 30 samples in this project.
- There are 0 comparisons in this project.

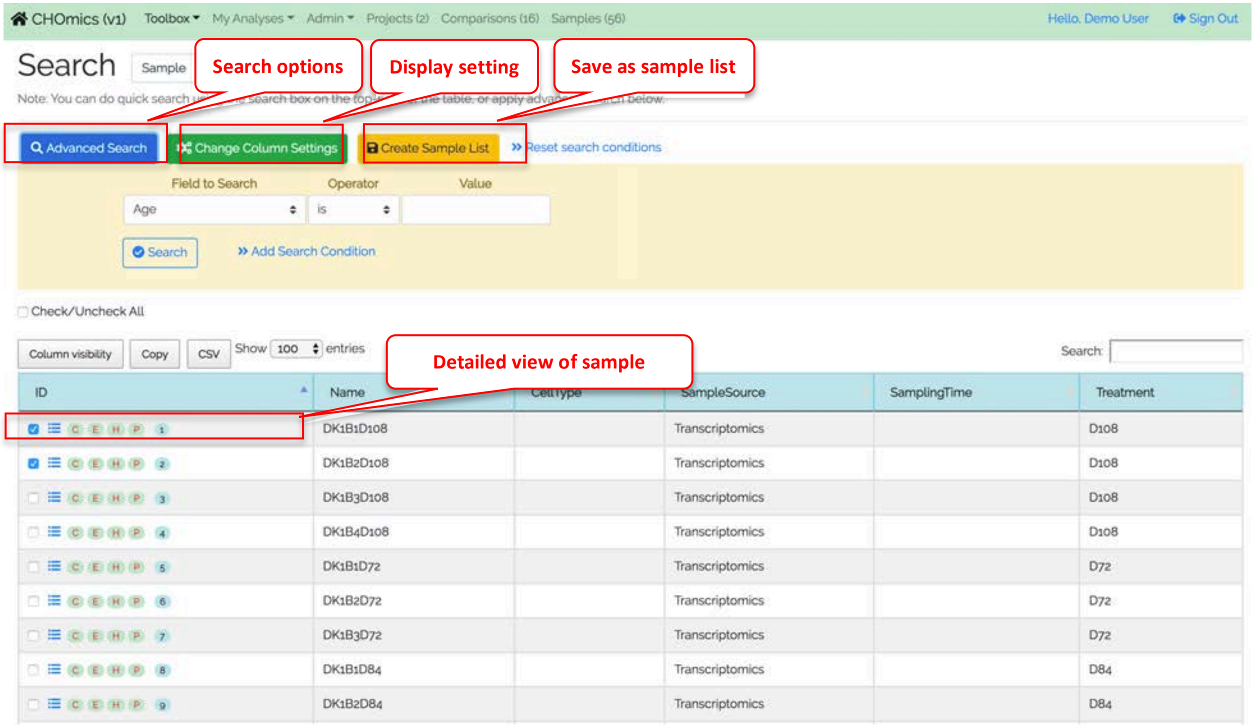
RNA-Seq Data - Test1

Created on 2019-10-23
- There are 26 samples in this project.
- There are 4 comparisons in this project.

Each project mainly consists of samples including both meta information and omics profiling, and comparisons showing the statistical differences among samples.

1.4 Samples

A project may include many samples which can be searched by the ‘Sample’ in top menu bar. Each sample has its own properties including Species, CellType, DiseaseState,etc (details available by clicking the ☰ button on the left ends).



Search

Search options

Display setting

Save as sample list

Note: You can do quick search using the search box on the top-right of the table, or apply advanced search below.

Field to Search: Age
Operator: is
Value:

□ Check/Uncheck All

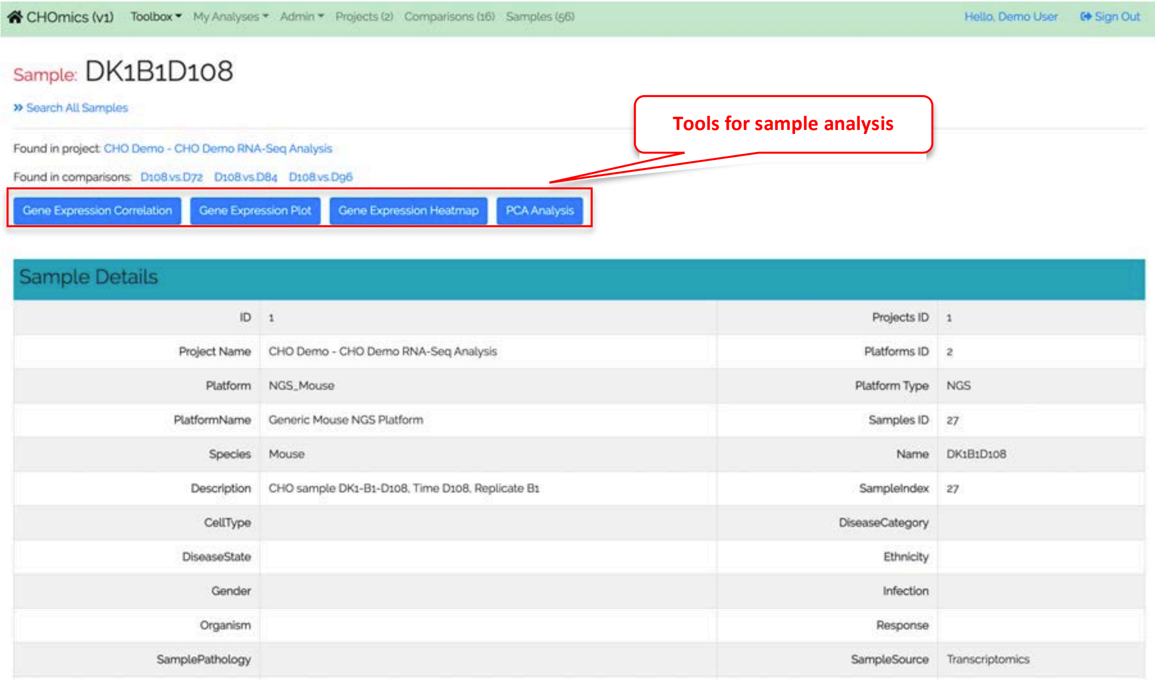
Show 100 entries

Detailed view of sample

ID	Name	CellType	SampleSource	SamplingTime	Treatment
☑ ☰ C E H P 1	DK1B1D108		Transcriptomics		D108
☑ ☰ C E H P 2	DK1B2D108		Transcriptomics		D108
□ ☰ C E H P 3	DK1B3D108		Transcriptomics		D108
□ ☰ C E H P 4	DK1B4D108		Transcriptomics		D108
□ ☰ C E H P 5	DK1B1D72		Transcriptomics		D72
□ ☰ C E H P 6	DK1B2D72		Transcriptomics		D72
□ ☰ C E H P 7	DK1B3D72		Transcriptomics		D72
□ ☰ C E H P 8	DK1B1D84		Transcriptomics		D84
□ ☰ C E H P 9	DK1B2D84		Transcriptomics		D84

To change columns displayed in the table, using the table settings (green button). Users can also select the samples to save them into the sample list. Samples from the list can be loaded to other analysis or visualization stools like heatmap.

Each sample has a gene expression profile. In CHOmics, there are multiple ways to analyze and visualize the samples including: correlation tool (noted by ‘C’), gene expression plot(noted by ‘E’), expression heatmap (noted by ‘H’), and PCA analysis (noted by ‘P’).



Sample: DK1B1D108

Found in project: CHO Demo - CHO Demo RNA-Seq Analysis

Found in comparisons: D108.vs.D72 D108.vs.D84 D108.vs.D96

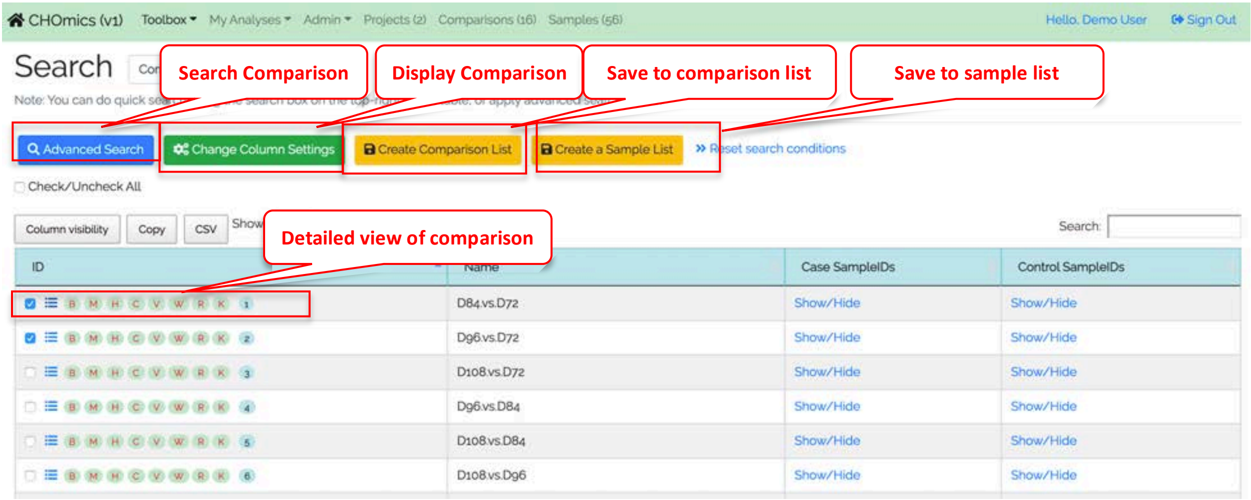
Sample Details

ID	1	Projects ID	1
Project Name	CHO Demo - CHO Demo RNA-Seq Analysis	Platforms ID	2
Platform	NGS_Mouse	Platform Type	NGS
PlatformName	Generic Mouse NGS Platform	Samples ID	27
Species	Mouse	Name	DK1B1D108
Description	CHO sample DK1-B1-D108, Time D108, Replicate B1	SampleIndex	27
CellType		DiseaseCategory	
DiseaseState		Ethnicity	
Gender		Infection	
Organism		Response	
SamplePathology		SampleSource	Transcriptomics

1.5 Comparisons

Comparison is defined by the comparative analysis between two groups of samples including differential gene analysis and pathway enrichment analysis.

There are a lot of meta data available for each comparison. See the dashboard for an overview of key categories, and the detailed description of each comparison has the full information.



Search

Note: You can do quick se… search box on the top-right… or apply advanced search

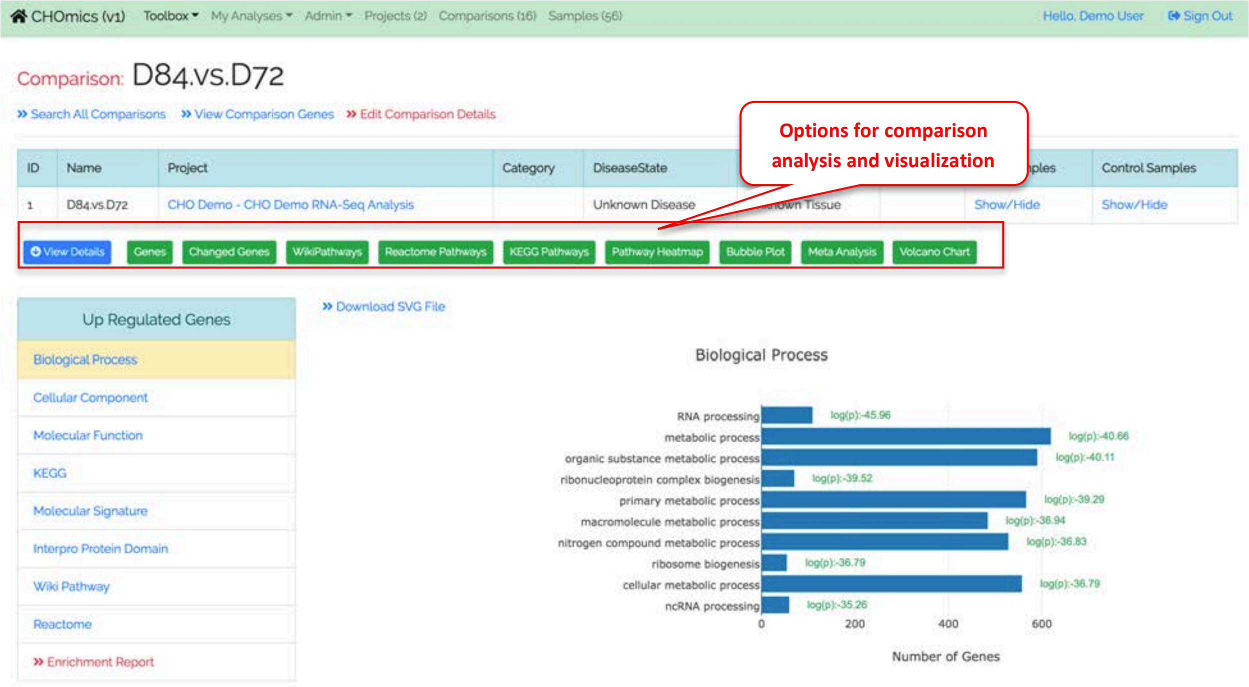
□ Check/Uncheck All

Search:

ID	Name	Case SampleIDs	Control SampleIDs
☑ B M H C V W R K 1	D84.vs.D72	Show/Hide	Show/Hide
☑ B M H C V W R K 2	D96.vs.D72	Show/Hide	Show/Hide
□ B M H C V W R K 3	D108.vs.D72	Show/Hide	Show/Hide
□ B M H C V W R K 4	D96.vs.D84	Show/Hide	Show/Hide
□ B M H C V W R K 5	D108.vs.D84	Show/Hide	Show/Hide
□ B M H C V W R K 6	D108.vs.D96	Show/Hide	Show/Hide

The selected comparisons can be saved to the comparison list (yellow button) for easy loading into the plotting tools.

Several options on each comparison for complicated visualization and analysis are also listed including: bubble plot of gene expressions(noted by ‘B’), meta analysis (noted by ‘M’), pathway heamap plot(noted by ‘H’), significant changes genes (noted by ‘C’), volcano plot(noted by ‘V’), Wikipathway mapping(noted by ‘W’), and Rectome and KEGG pathway mapping (noted by ‘R’ and ‘K’ respectively).



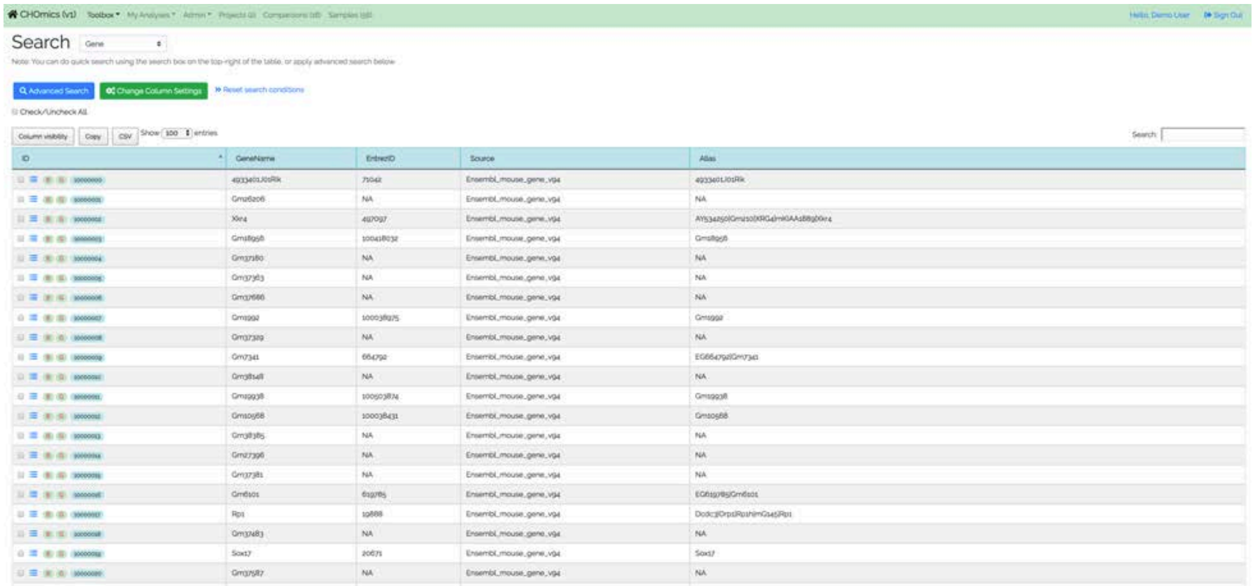
Comparison: D84.vs.D72

» Search All Comparisons » View Comparison Genes » Edit Comparison Details

ID	Name	Project	Category	DiseaseState		…ples	Control Samples
1	D84.vs.D72	CHO Demo - CHO Demo RNA-Seq Analysis		Unknown Disease	…own Tissue	Show/Hide	Show/Hide

1.6 Genes

The genome-wide gene expression values were detected in each sample using RNA-Seq or microarrays. All the human genes that have expression values are listed in gene table. The gene annotation from difference platforms were all mapped to NCBI gene ID (EntrezID) for consistence across platforms.



Note: You can do quick search using the search box on the top-right of the table, or apply advanced search below

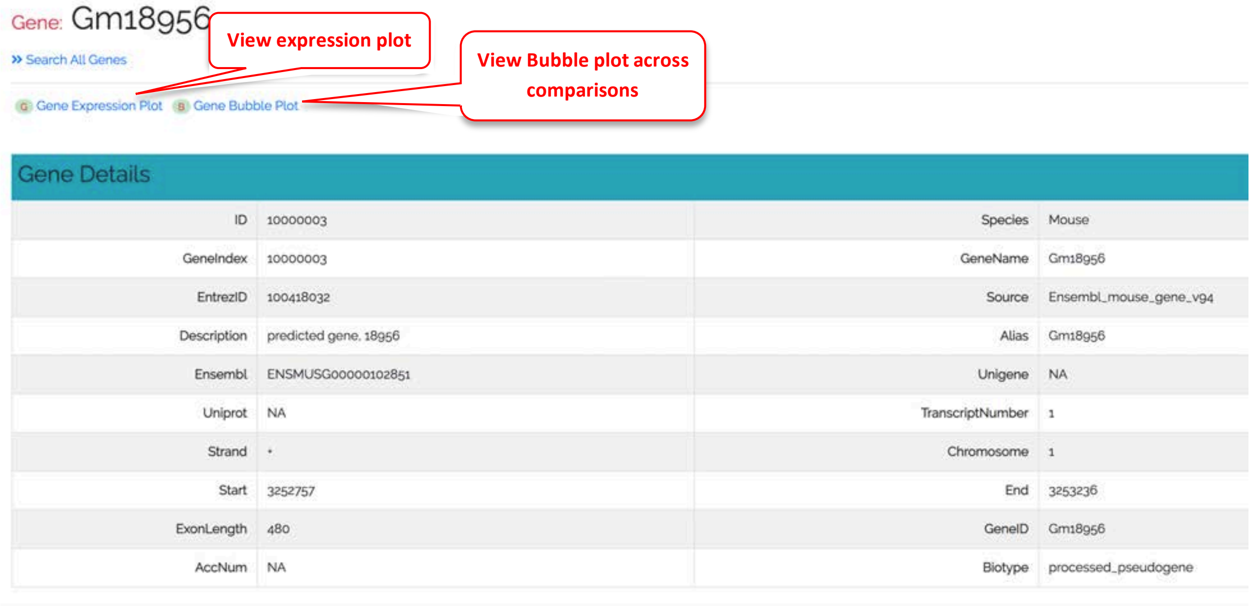
ID	GeneName	EntrezID	Source	Alias
□ ☰ 10000000	4933401J01Rik	71042	Ensembl_mouse_gene_v94	4933401J01Rik
□ ☰ 10000001	Gm26206	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000002	Xkr4	497097	Ensembl_mouse_gene_v94	AY534250|Gm210|XRG4|mKIAA1889|Xkr4
□ ☰ 10000003	Gm18956	100418032	Ensembl_mouse_gene_v94	Gm18956
□ ☰ 10000004	Gm37180	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000005	Gm37363	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000006	Gm37686	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000007	Gm1992	100038975	Ensembl_mouse_gene_v94	Gm1992
□ ☰ 10000008	Gm37329	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000009	Gm7341	664792	Ensembl_mouse_gene_v94	EG664792|Gm7341
□ ☰ 10000010	Gm38148	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000011	Gm19938	100503874	Ensembl_mouse_gene_v94	Gm19938
□ ☰ 10000012	Gm10568	100038431	Ensembl_mouse_gene_v94	Gm10568
□ ☰ 10000013	Gm38385	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000014	Gm27396	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000015	Gm37381	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000016	Gm6101	619785	Ensembl_mouse_gene_v94	EG619785|Gm6101
□ ☰ 10000017	Rp1	19888	Ensembl_mouse_gene_v94	Dcdc3|Orp1|Rp1h|mG145|Rp1
□ ☰ 10000018	Gm37483	NA	Ensembl_mouse_gene_v94	NA
□ ☰ 10000019	Sox17	20671	Ensembl_mouse_gene_v94	Sox17
□ ☰ 10000020	Gm37587	NA	Ensembl_mouse_gene_v94	NA

To find a gene, you can use gene symbol, gene description, gene alias, NCBI gene ID, Ensembl gene ID or Uniprot ID.

For some common genes, the symbols used in publications are often not the official symbol, and you can try search alias field. For example, TP53 is often referred to as P53 in publication. You need to search P53 in alias or tumor protein p53 in description to find it if you don't know its official symbol.

The NCBI Gene search https://www.ncbi.nlm.nih.gov/gene is a good source to get official gene symbols and IDs.

You can view full details of a gene by clicking the ☰ button .



Gene Details

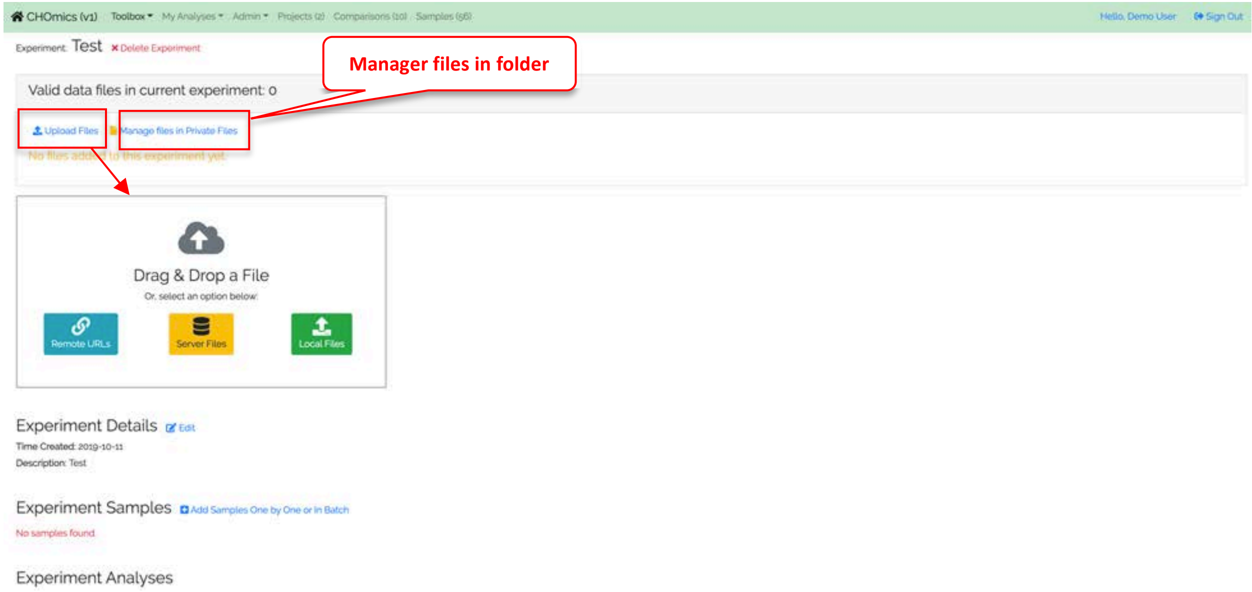
ID	10000003	Species	Mouse
GeneIndex	10000003	GeneName	Gm18956
EntrezID	100418032	Source	Ensembl_mouse_gene_v94
Description	predicted gene, 18956	Alias	Gm18956
Ensembl	ENSMUSG00000102851	Unigene	NA
Uniprot	NA	TranscriptNumber	1
Strand	-	Chromosome	1
Start	3252757	End	3253236
ExonLength	480	GeneID	Gm18956
AccNum	NA	Biotype	processed_pseudogene

From gene details, you can access RNA-Seq data in a box plot, or view all comparisons including this gene in a bubble plot.

2 Data Input

2.1 Upload fastq files to experiment

After the experiment is created, users can upload fastq or fastq.gz files through remote URLs, server files or local files. The files are uploaded to the private folder named 'Experiments' automatically.



Experiment: Test

Valid data files in current experiment: 0

No files added to this experiment yet.

Experiment Details

Time Created: 2019-10-11

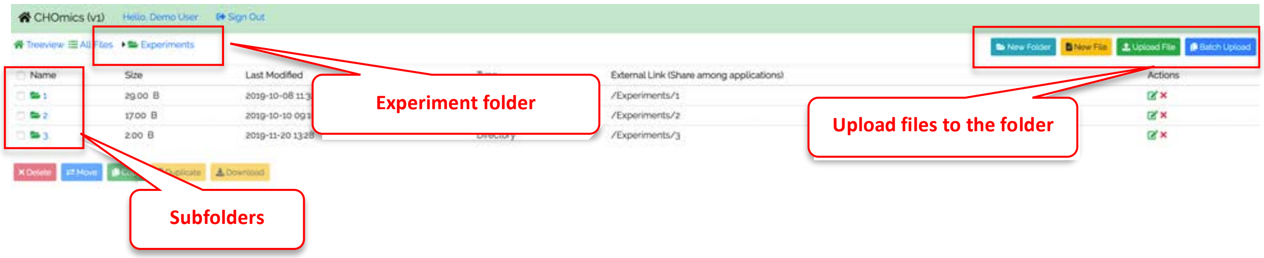
Description: Test

Experiment Samples

No samples found

Experiment Analyses

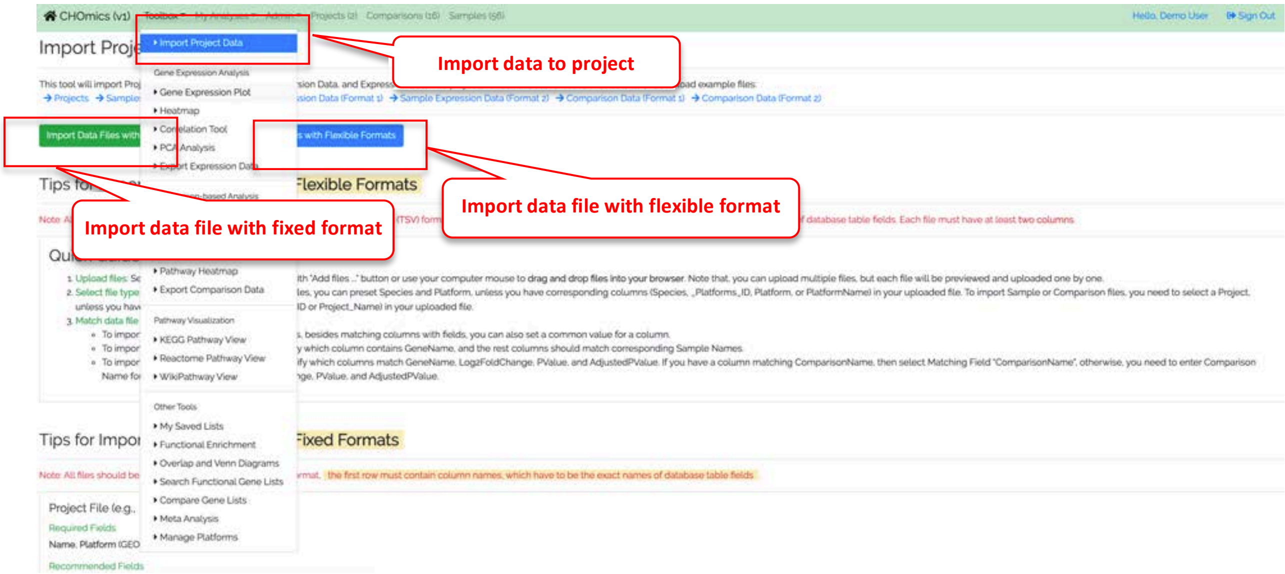
In the folder ‘Experiments’, there may be multiple subfolders corresponding to different experiments. Users can easily modify the folder or upload new files to the folder.



Name	Size	Last Modified	Type	External Link (Share among applications)	Actions
1	29.00 B	2019-10-08 11:3		/Experiments/1	
2	17.00 B	2019-10-10 09:1		/Experiments/2	
3	2.00 B	2019-11-20 13:28	Directory	/Experiments/3	

2.2 Upload data file to project

Besides raw RNA sequencing data (fastq files), CHOmics also allow the input of other types of data to start a project, including meta data (i.e,project and samples), expression data, and summary data. Those data should be uploaded in comma separated values (CSV) or tab separated values (TSV) with either fixed or flexible format.

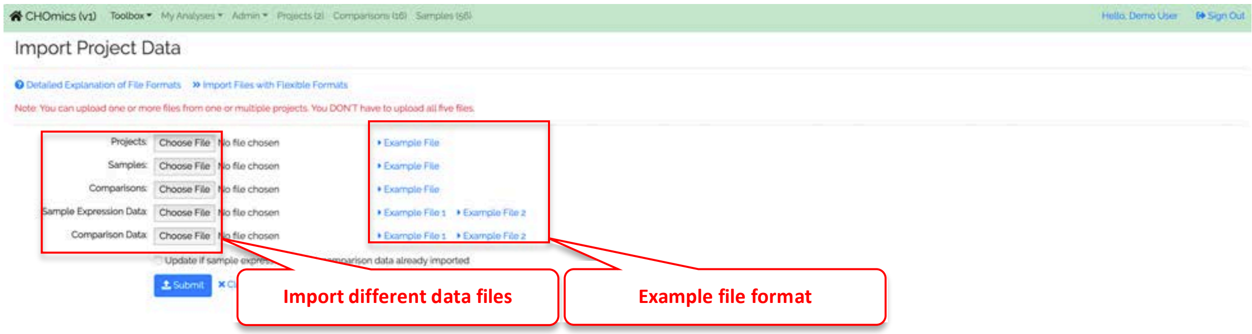


Project file can be uploaded to create a new project. The project file contains some required information such as Name, Platform and other optional fields such as Disease, Description etc.

Sample file can be uploaded to register samples for a project. The sample file contains required information such as Name and Project_Name and optional fields such as Description, Tissue, DiseaseState, SampleSource, Gender, etc.

Expression file can be uploaded with quantified expression measure at gene level. The expression could be transcriptomics, proteomics or other gene-level counts. The file is required to contain GeneName, SampleName, Value.

Comparison file and Comparison data file are used to upload summary results for statistical comparison test applied externally. Comparison file needs to contain the Project_Name, Case_SampleIDs, and Control_SampleIDs while comparison data file contains statistical results such as GeneName, ComparisonName, Log2FoldChange, PValue, Adjusted PValue for each comparison.


Import Project Data

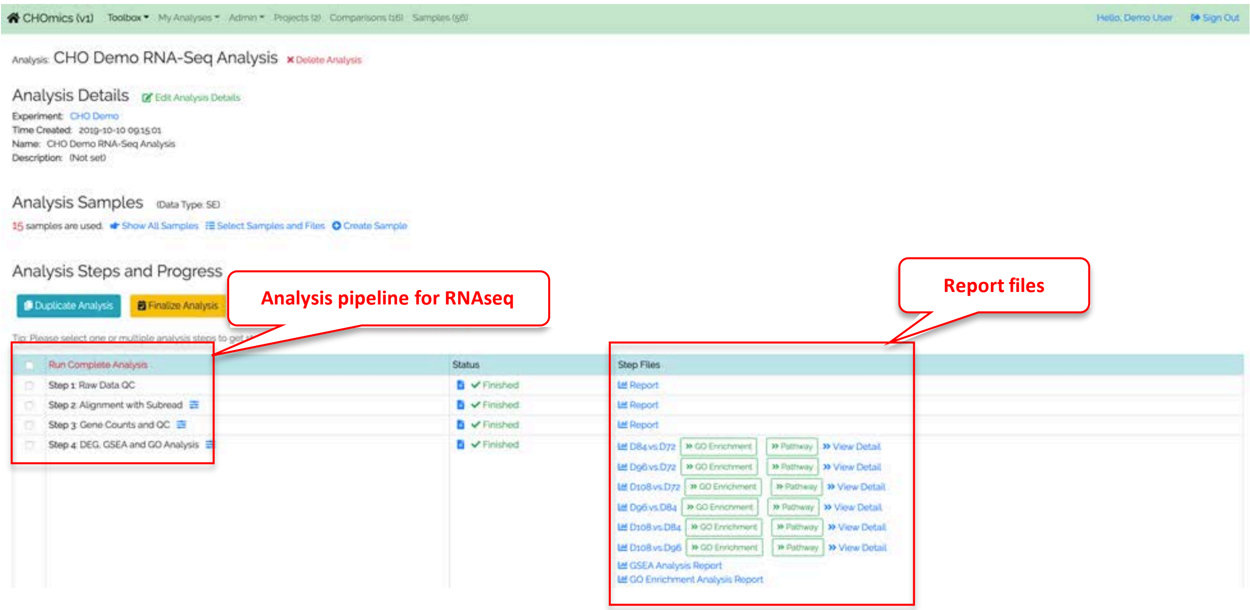
Note: You can upload one or more files from one or multiple projects. You DON'T have to upload all five files.

CHOmics (v1)
Toolbox
My Analyses
Admin
Projects (2)
Comparisons (16)
Samples (56)
Hello, Demo User
Sign Out
Detailed Explanation of File Formats
Import Files with Flexible Formats
Projects: Choose File No file chosen
Samples: Choose File No file chosen
Comparisons: Choose File No file chosen
Sample Expression Data: Choose File No file chosen
Comparison Data: Choose File No file chosen
Example File
Example File
Example File
Example File 1
Example File 2
Example File 1
Example File 2
Update if sample expres
mparison data already imported
Submit
Import different data files
Example file format

3 Data Analysis

3.1 RNAseq analysis pipeline

After fastq files are uploaded to the experiment by following the Section 2.1, users can start the analysis by applying the built-in pipeline mainly including: Raw Data QC (quality control), Alignment, Gene Counts and QC, and DEG, GSEA and GO analysis. After the analysis is completed, the results can be exported into a project for visualization.



Analysis: CHO Demo RNA-Seq Analysis

Analysis Details

Experiment: CHO Demo
Time Created: 2019-10-10 09:15:01
Name: CHO Demo RNA-Seq Analysis
Description: (Not set)

Analysis Samples (Data Type: SE)

15 samples are used.

Analysis Steps and Progress

Tip: Please select one or multiple analysis steps to get st

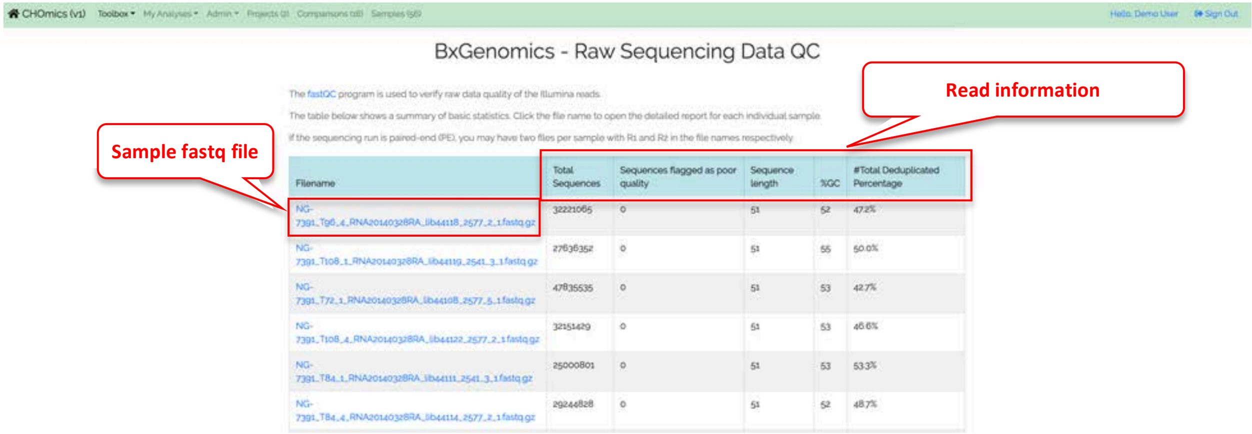
□ Run Complete Analysis	Status	Step Files
□ Step 1: Raw Data QC	✓ Finished	Report
□ Step 2: Alignment with Subread	✓ Finished	Report
□ Step 3: Gene Counts and QC	✓ Finished	Report
□ Step 4: DEG, GSEA and GO Analysis	✓ Finished	D84 vs D72 » GO Enrichment » Pathway » View Detail D96 vs D72 » GO Enrichment » Pathway » View Detail D108 vs D72 » GO Enrichment » Pathway » View Detail D96 vs D84 » GO Enrichment » Pathway » View Detail D108 vs D84 » GO Enrichment » Pathway » View Detail D108 vs D96 » GO Enrichment » Pathway » View Detail GSEA Analysis Report GO Enrichment Analysis Report

Analysis pipeline for RNAseq

Report files

After completion of each step, a report is generated for summarizing the metrics in each step to quantify raw data QC, alignment with Subread method, and gene count distribution and sample/gene count QC, respectively.

In the report for raw data QC, all fastq files are verified in quality by software fastQC. Sequencing read information and quality control metrics are summarized for each individual fastq file.



BxGenomics - Raw Sequencing Data QC

The fastQC program is used to verify raw data quality of the Illumina reads.

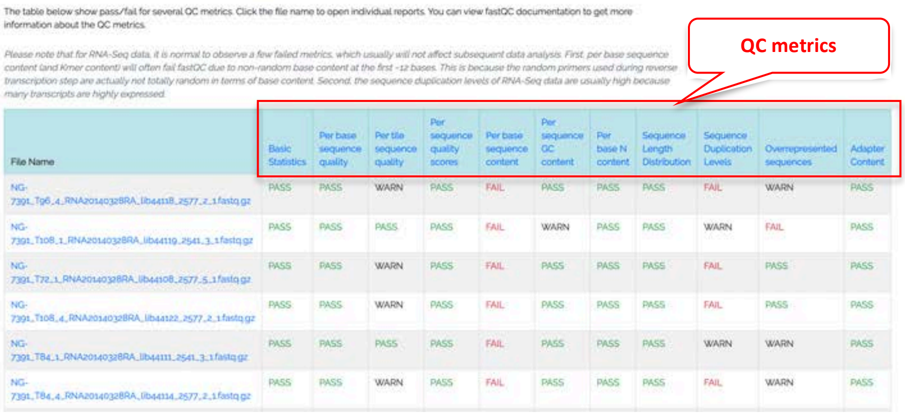
The table below shows a summary of basic statistics. Click the file name to open the detailed report for each individual sample.

If the sequencing run is paired-end (PE), you may have two files per sample with R1 and R2 in the file names respectively.

Filename	Total Sequences	Sequences flagged as poor quality	Sequence length	%GC	#Total Deduplicated Percentage
NG-7391_T96_4_RNA20140328RA_lib44118_2577_2_1.fastq.gz	32221065	0	51	52	47.2%
NG-7391_T108_1_RNA20140328RA_lib44119_2541_3_1.fastq.gz	27636352	0	51	55	50.0%
NG-7391_T72_1_RNA20140328RA_lib44108_2577_5_1.fastq.gz	47835535	0	51	53	42.7%
NG-7391_T108_4_RNA20140328RA_lib44122_2577_2_1.fastq.gz	32151429	0	51	53	46.6%
NG-7391_T84_1_RNA20140328RA_lib44111_2541_3_1.fastq.gz	25000801	0	51	53	53.3%
NG-7391_T84_4_RNA20140328RA_lib44114_2577_2_1.fastq.gz	29244828	0	51	52	48.7%

Sample fastq file

Read information



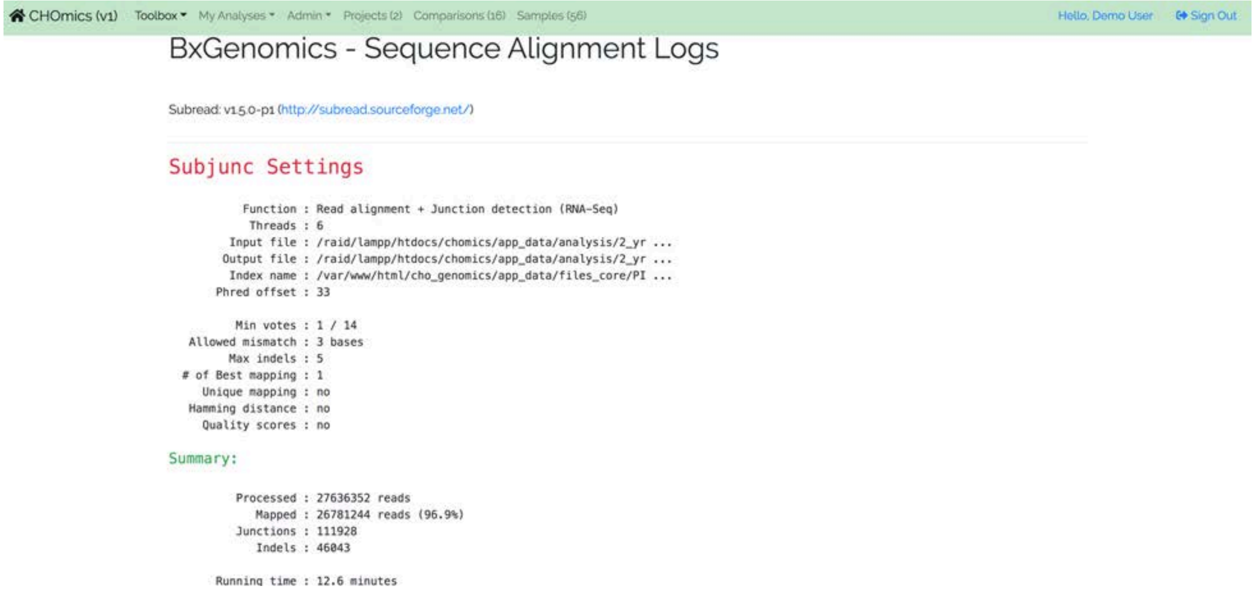
The table below show pass/fail for several QC metrics. Click the file name to open individual reports. You can view fastQC documentation to get more information about the QC metrics.

Please note that for RNA-Seq data, it is normal to observe a few failed metrics, which usually will not affect subsequent data analysis. First, per base sequence content (and Kmer content) will often fail fastQC due to non-random base content at the first ~12 bases. This is because the random primers used during reverse transcription step are actually not totally random in terms of base content. Second, the sequence duplication levels of RNA-Seq data are usually high because many transcripts are highly expressed.

QC metrics

File Name	Basic Statistics	Per base sequence quality	Per tile sequence quality	Per sequence quality scores	Per base sequence content	Per sequence GC content	Per base N content	Sequence Length Distribution	Sequence Duplication Levels	Overrepresented sequences	Adapter Content
NG-7391_T96_4_RNA20140328RA_lib44118_2577_2_1.fastq.gz	PASS	PASS	WARN	PASS	FAIL	PASS	PASS	PASS	FAIL	WARN	PASS
NG-7391_T108_1_RNA20140328RA_lib44119_2541_3_1.fastq.gz	PASS	PASS	PASS	PASS	FAIL	WARN	PASS	PASS	WARN	FAIL	PASS
NG-7391_T72_1_RNA20140328RA_lib44108_2577_5_1.fastq.gz	PASS	PASS	WARN	PASS	FAIL	PASS	PASS	PASS	FAIL	PASS	PASS
NG-7391_T108_4_RNA20140328RA_lib44122_2577_2_1.fastq.gz	PASS	PASS	WARN	PASS	FAIL	PASS	PASS	PASS	FAIL	PASS	PASS
NG-7391_T84_1_RNA20140328RA_lib44111_2541_3_1.fastq.gz	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	PASS	WARN	WARN	PASS
NG-7391_T84_4_RNA20140328RA_lib44114_2577_2_1.fastq.gz	PASS	PASS	WARN	PASS	FAIL	PASS	PASS	PASS	FAIL	WARN	PASS

In the report for alignment, parameter setting and quality metrics (e.g, mapped, junctions,etc) for alignment are listed for each fastq file.



BxGenomics - Sequence Alignment Logs

Subread: v1.5.0-p1 (http://subread.sourceforge.net/)

Subjunc Settings

```
          Function : Read alignment + Junction detection (RNA-Seq)
           Threads : 6
        Input file : /raid/lampp/htdocs/chomics/app_data/analysis/2_yr ...
       Output file : /raid/lampp/htdocs/chomics/app_data/analysis/2_yr ...
        Index name : /var/www/html/cho_genomics/app_data/files_core/PI ...
      Phred offset : 33

         Min votes : 1 / 14
  Allowed mismatch : 3 bases
        Max indels : 5
 # of Best mapping : 1
    Unique mapping : no
  Hamming distance : no
    Quality scores : no
```

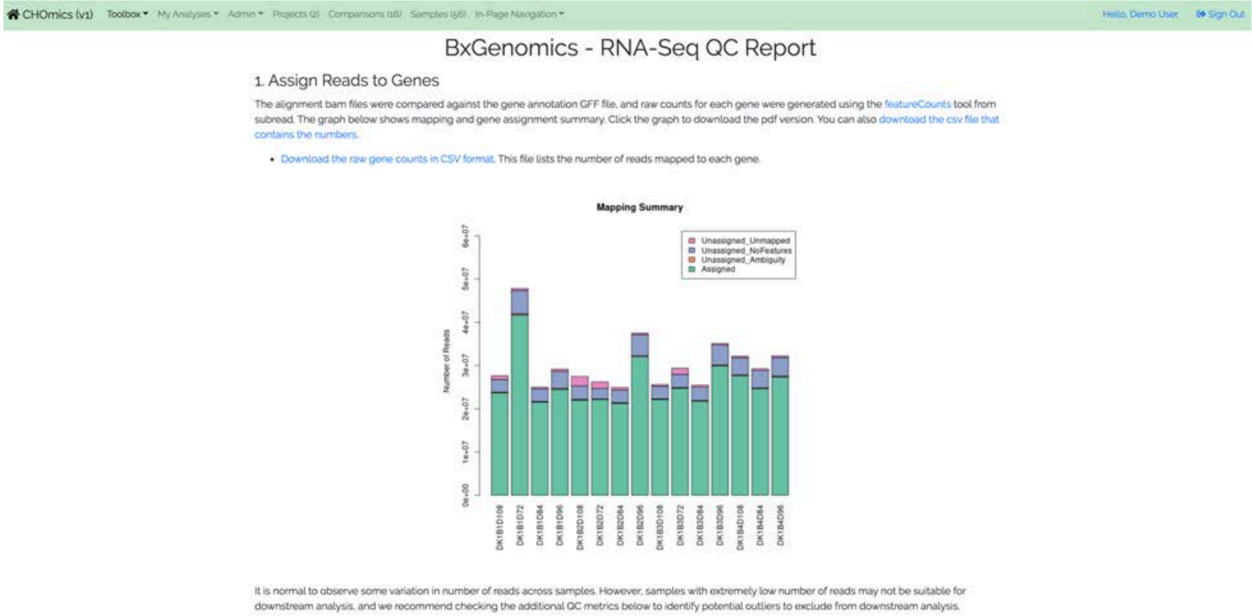
Summary:

```
         Processed : 27636352 reads
            Mapped : 26781244 reads (96.9%)
         Junctions : 111928
            Indels : 46043

      Running time : 12.6 minutes
```

In the report for Gene Counts and QC step, several metrics have been calculated and plotted for comprehensive evaluation of genes and samples, including: reads mapping to genes, distribution of detected genes, percentage of reads for highly expressed genes, normalization and boxplot of gene expression, sample grouping and clustering, sample correlation and outlier detection.

‘1. Assign reads to genes’ plots the mapping summary of reads to genes, showing the percentage of reads assigned to genes or unassigned due to unmapping, no features or ambiguous mapping.



BxGenomics - RNA-Seq QC Report

1. Assign Reads to Genes

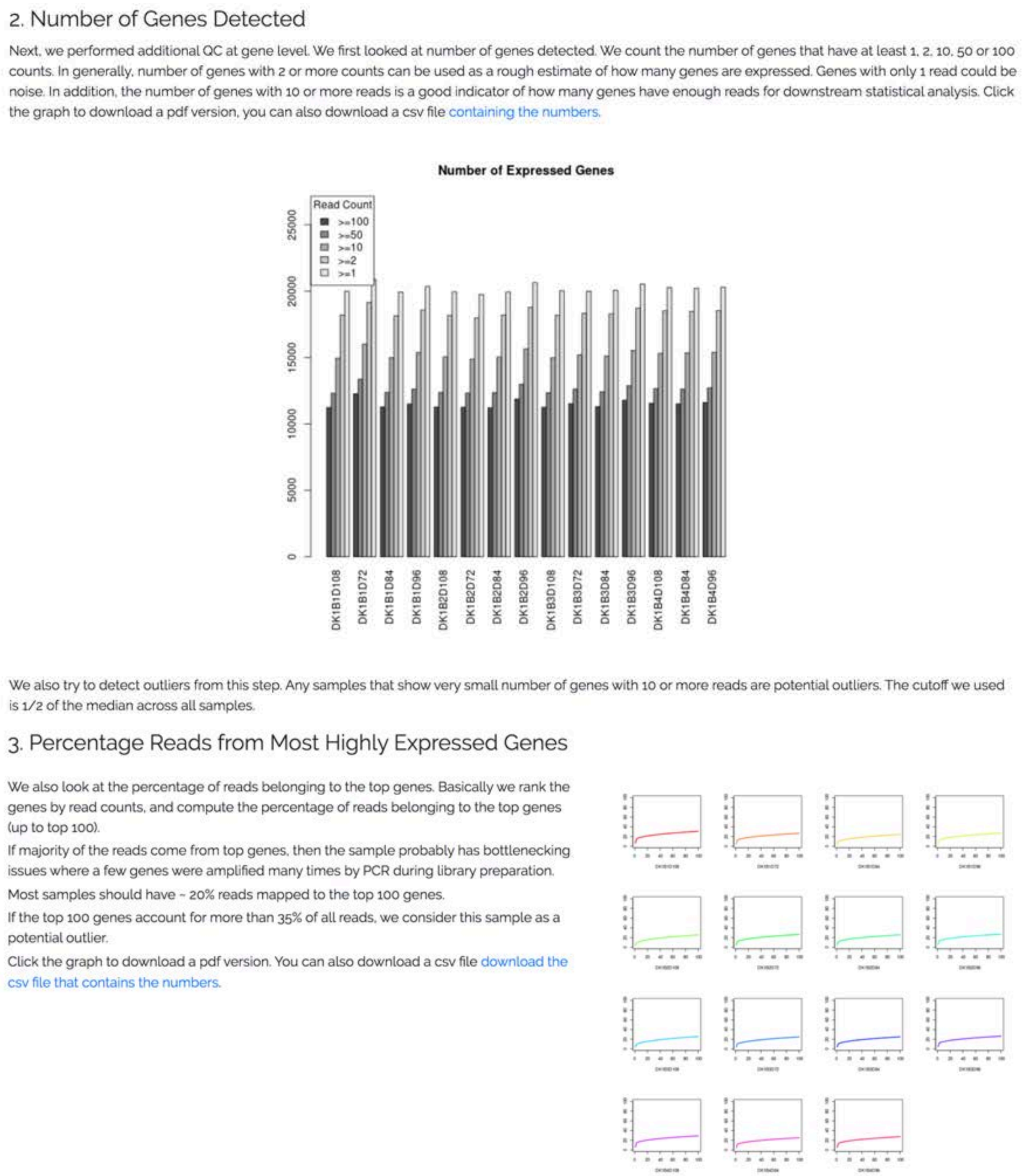
The alignment bam files were compared against the gene annotation GFF file, and raw counts for each gene were generated using the featureCounts tool from subread. The graph below shows mapping and gene assignment summary. Click the graph to download the pdf version. You can also download the csv file that contains the numbers.

- Download the raw gene counts in CSV format. This file lists the number of reads mapped to each gene.

It is normal to observe some variation in number of reads across samples. However, samples with extremely low number of reads may not be suitable for downstream analysis, and we recommend checking the additional QC metrics below to identify potential outliers to exclude from downstream analysis.

‘2. Number of Genes Detected’ plots the number of expressed genes with read count from intervals of >=1, >=2, >=10, >=50 and >=100.

‘3. Percentage Reads from Most Highly Expressed Genes’ plots the percentage of reads mapped to the top expressed genes (up to 100 genes).



2. Number of Genes Detected

Next, we performed additional QC at gene level. We first looked at number of genes detected. We count the number of genes that have at least 1, 2, 10, 50 or 100 counts. In generally, number of genes with 2 or more counts can be used as a rough estimate of how many genes are expressed. Genes with only 1 read could be noise. In addition, the number of genes with 10 or more reads is a good indicator of how many genes have enough reads for downstream statistical analysis. Click the graph to download a pdf version, you can also download a csv file containing the numbers.

We also try to detect outliers from this step. Any samples that show very small number of genes with 10 or more reads are potential outliers. The cutoff we used is 1/2 of the median across all samples.

3. Percentage Reads from Most Highly Expressed Genes

We also look at the percentage of reads belonging to the top genes. Basically we rank the genes by read counts, and compute the percentage of reads belonging to the top genes (up to top 100).

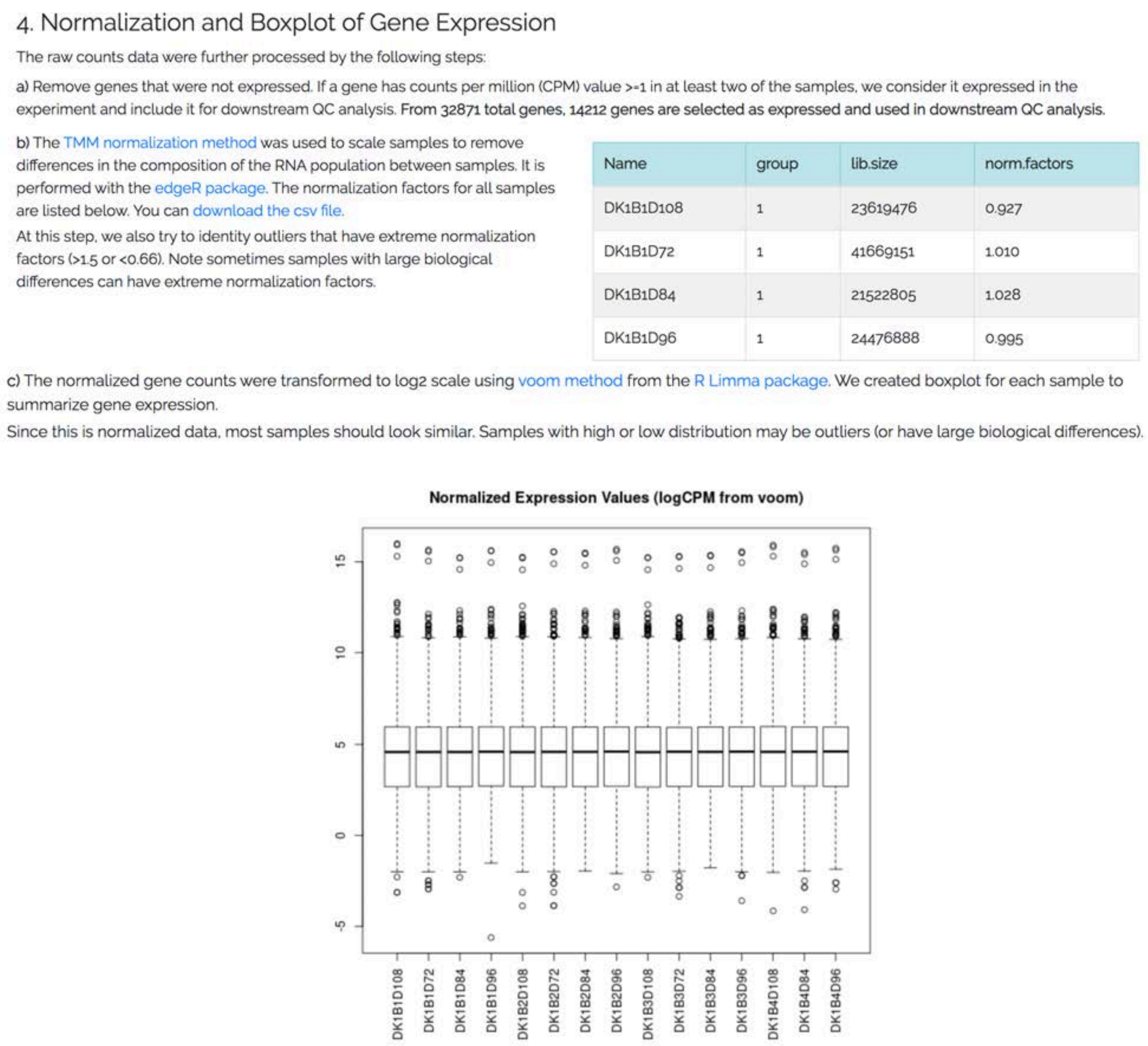
If majority of the reads come from top genes, then the sample probably has bottlenecking issues where a few genes were amplified many times by PCR during library preparation.

Most samples should have ~ 20% reads mapped to the top 100 genes.

If the top 100 genes account for more than 35% of all reads, we consider this sample as a potential outlier.

Click the graph to download a pdf version. You can also download a csv file download the csv file that contains the numbers.

‘4. Normalization and Boxplot of Gene Expression’ evaluates gene expression after normalization by TMM method and then draws boxplot of normalized expression (logCPM: log of counts per million reads) after log2 transformation.



4. Normalization and Boxplot of Gene Expression

The raw counts data were further processed by the following steps:

a) Remove genes that were not expressed. If a gene has counts per million (CPM) value >=1 in at least two of the samples, we consider it expressed in the experiment and include it for downstream QC analysis. **From 32871 total genes, 14212 genes are selected as expressed and used in downstream QC analysis.**

b) The TMM normalization method was used to scale samples to remove differences in the composition of the RNA population between samples. It is performed with the edgeR package. The normalization factors for all samples are listed below. You can download the csv file.

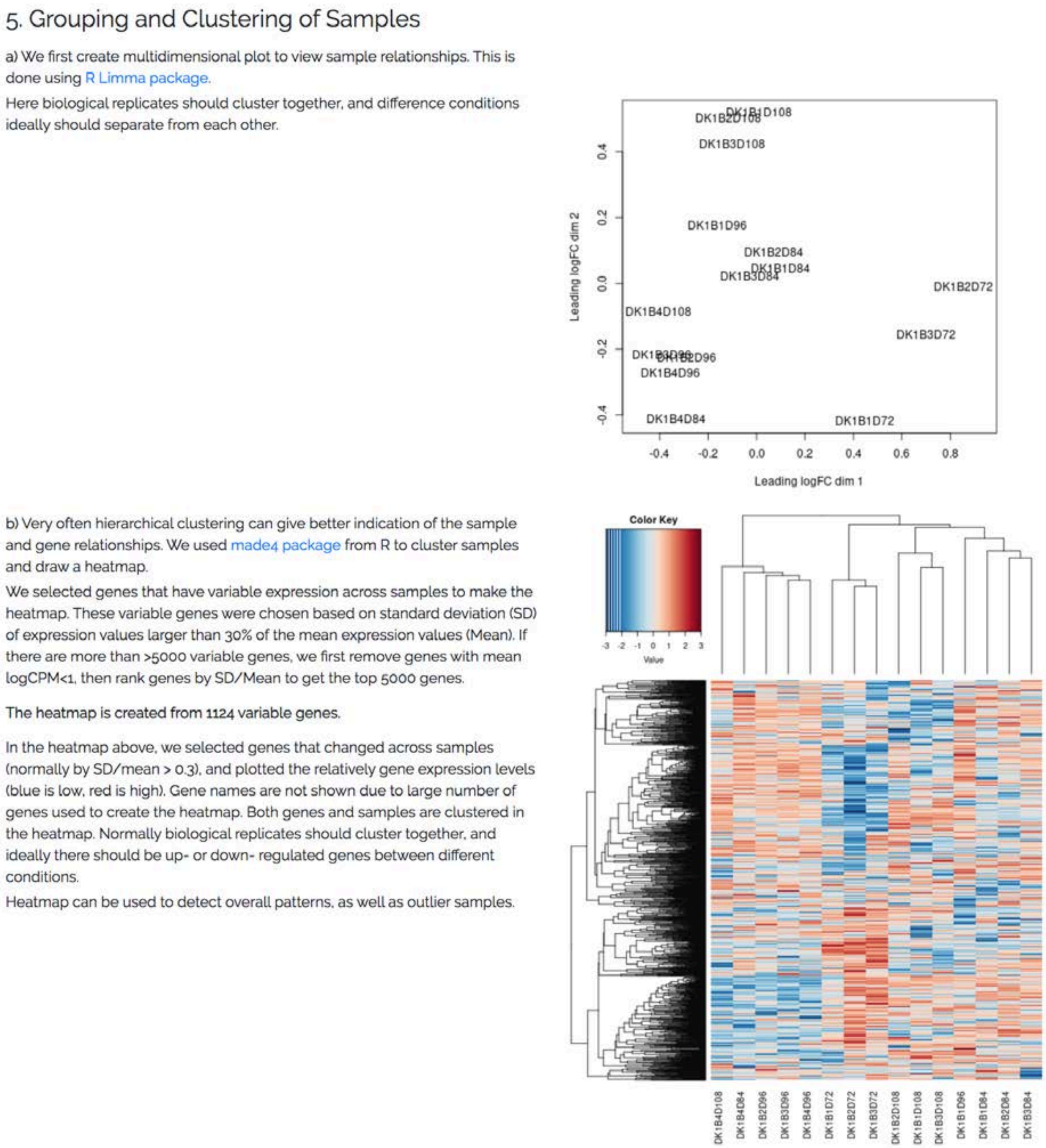
At this step, we also try to identity outliers that have extreme normalization factors (>1.5 or <0.66). Note sometimes samples with large biological differences can have extreme normalization factors.

Name	group	lib.size	norm.factors
DK1B1D108	1	23619476	0.927
DK1B1D72	1	41669151	1.010
DK1B1D84	1	21522805	1.028
DK1B1D96	1	24476888	0.995

c) The normalized gene counts were transformed to log2 scale using voom method from the R Limma package. We created boxplot for each sample to summarize gene expression.

Since this is normalized data, most samples should look similar. Samples with high or low distribution may be outliers (or have large biological differences).

‘5 Grouping and Clustering of Samples’ plots the relationship among samples by multidimensional scale. Samples are clustered by hierarchical clustering method based on the expression of top genes with large variation(SD/mean>0.3).



5. Grouping and Clustering of Samples

a) We first create multidimensional plot to view sample relationships. This is done using R Limma package.

Here biological replicates should cluster together, and difference conditions ideally should separate from each other.

b) Very often hierarchical clustering can give better indication of the sample and gene relationships. We used made4 package from R to cluster samples and draw a heatmap.

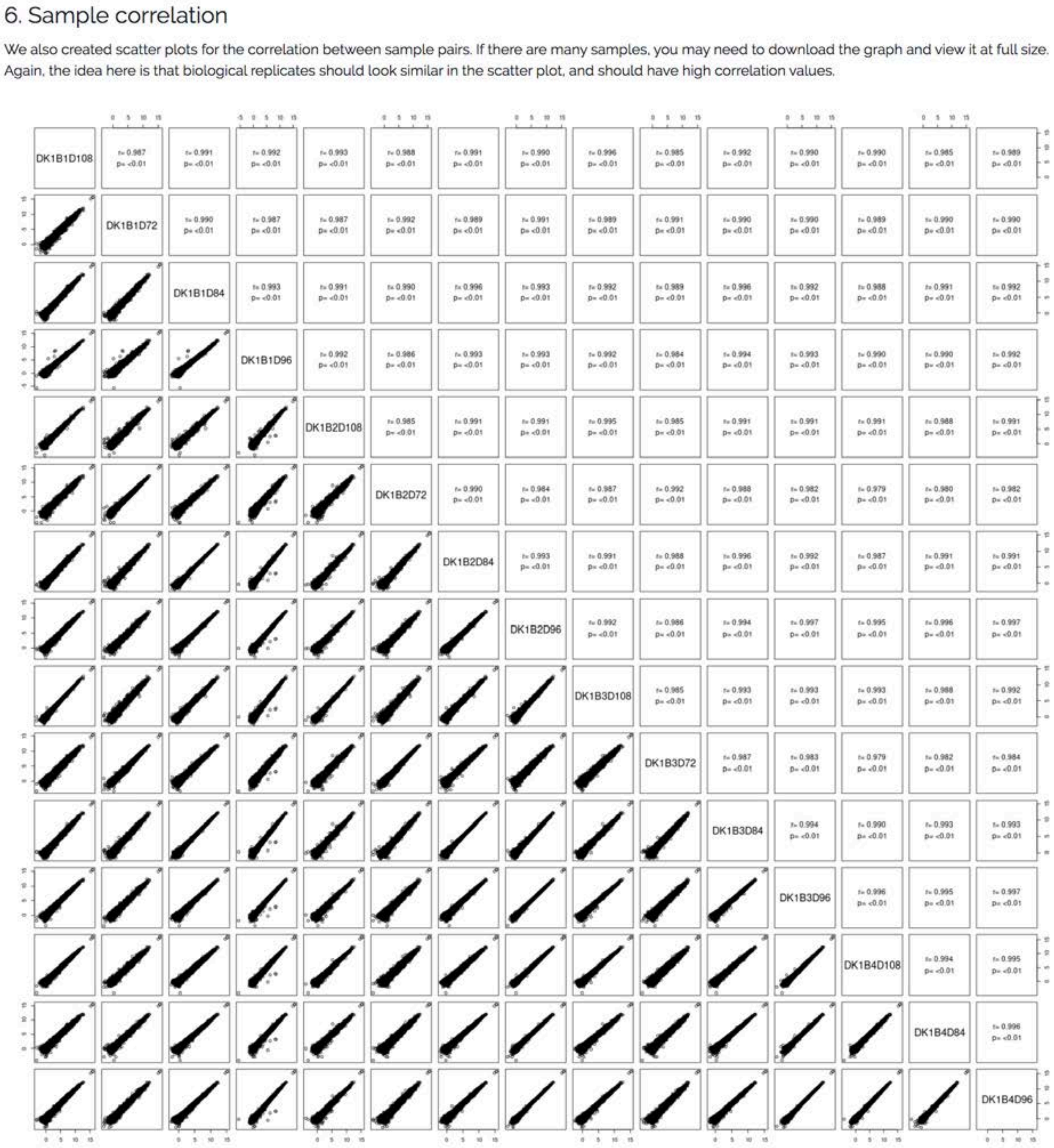
We selected genes that have variable expression across samples to make the heatmap. These variable genes were chosen based on standard deviation (SD) of expression values larger than 30% of the mean expression values (Mean). If there are more than >5000 variable genes, we first remove genes with mean logCPM<1, then rank genes by SD/Mean to get the top 5000 genes.

The heatmap is created from 1124 variable genes.

In the heatmap above, we selected genes that changed across samples (normally by SD/mean > 0.3), and plotted the relatively gene expression levels (blue is low, red is high). Gene names are not shown due to large number of genes used to create the heatmap. Both genes and samples are clustered in the heatmap. Normally biological replicates should cluster together, and ideally there should be up- or down- regulated genes between different conditions.

Heatmap can be used to detect overall patterns, as well as outlier samples.

‘6 Sample correlation’ creates scatter plots for the correlation between sample pairs. The idea is that biological replicates from the same group should look similar in the scatter plot, and should have high correlation values compared to the samples from other groups.



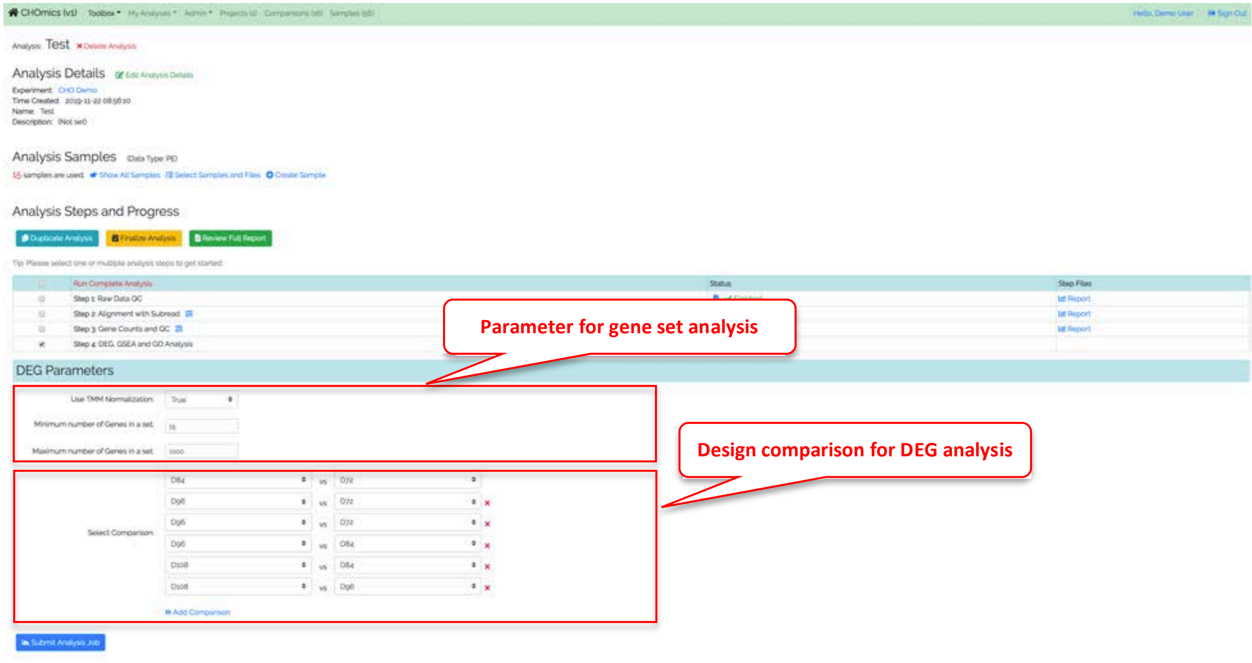
6. Sample correlation

We also created scatter plots for the correlation between sample pairs. If there are many samples, you may need to download the graph and view it at full size. Again, the idea here is that biological replicates should look similar in the scatter plot, and should have high correlation values.

3.2 DE, GSEA and GO analysis

After completing the first three steps for sample quality control and gene count readout, users can start statistical analysis as the last step of pipeline, including differential expression analysis (DEG), gene set enrichment analysis (GSEA) and gene ontology (GO) analysis.

DEG analysis is applied to compare gene expression between two groups, namely comparison. Users can design one or multiple comparisons for DEG analysis. In each comparison, differential expressed genes are identified by LIMMA model, followed by GSEA pathway analysis and GO enrichment analysis which explore the enrichment of DEGs in diverse pathways.



Analysis: Test ✖ Delete Analysis

Analysis Details Edit Analysis Details

Experiment: CHO Demo
Time Created: 2019-11-22 08:56:10
Name: Test
Description: (Not set)

Analysis Samples (Data Type: PE)

15 samples are used Show All Samples Select Samples and Files Create Sample

Analysis Steps and Progress

Tip: Please select one or multiple analysis steps to get started:

		Status	Step Files
□	Run Complete Analysis		
□	Step 1: Raw Data QC	✓ Finished	Report
□	Step 2: Alignment with Subread		Report
□	Step 3: Gene Counts and QC		Report
☑	Step 4: DEG, GSEA and GO Analysis		

DEG Parameters

Use TMM Normalization: True
Minimum number of Genes in a set: 15
Maximum number of Genes in a set: 1000

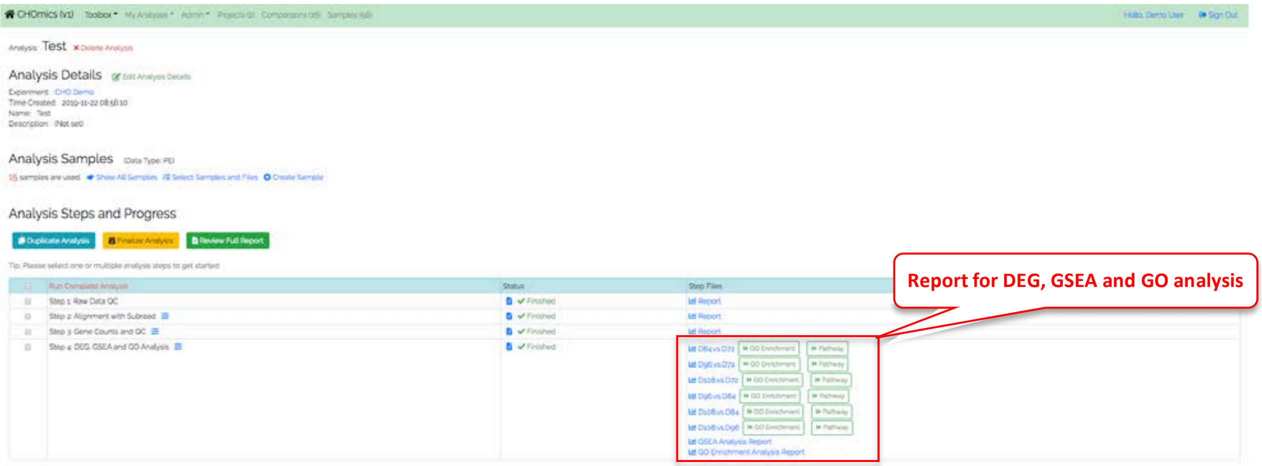
Select Comparison:

D84	vs	D72
D96	vs	D72
D96	vs	D72
D96	vs	D84
D108	vs	D84
D108	vs	D96

Parameter for gene set analysis

Design comparison for DEG analysis

The reports for DEG and pathway analysis are attached for each comparison after completion of analysis.



Analysis: Test ✖ Delete Analysis

Analysis Details Edit Analysis Details

Experiment: CHO Demo
Time Created: 2019-11-22 08:56:10
Name: Test
Description: (Not set)

Analysis Samples (Data Type: PE)

15 samples are used Show All Samples Select Samples and Files Create Sample

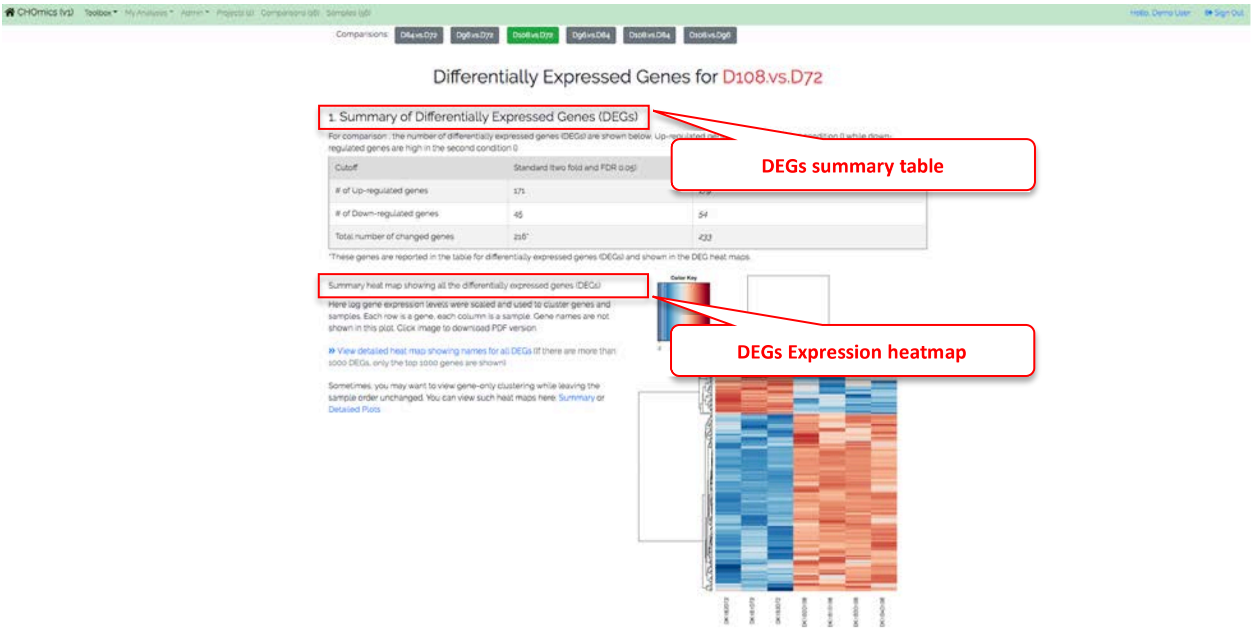
Analysis Steps and Progress

Tip: Please select one or multiple analysis steps to get started:

		Status	Step Files
□	Run Complete Analysis		
□	Step 1: Raw Data QC	✓ Finished	Report
□	Step 2: Alignment with Subread	✓ Finished	Report
□	Step 3: Gene Counts and QC	✓ Finished	Report
□	Step 4: DEG, GSEA and GO Analysis	✓ Finished	D84 vs D72 GO Enrichment Pathway D96 vs D72 GO Enrichment Pathway D108 vs D72 GO Enrichment Pathway D96 vs D84 GO Enrichment Pathway D108 vs D84 GO Enrichment Pathway D108 vs D96 GO Enrichment Pathway GSEA Analysis Report GO Enrichment Analysis Report

Report for DEG, GSEA and GO analysis

In the report for DEG analysis, the table summarizing DEGs with up- and down-regulation is listed along with a heatmap clustering the DEGs expression (up to top 1000 DEGs).



Differentially Expressed Genes for D108.vs.D72

1. Summary of Differentially Expressed Genes (DEGs)

For comparison : the number of differentially expressed genes (DEGs) are shown below. Up-regulated ge condition () while down-regulated genes are high in the second condition ()

Cutoff	Standard (two fold and FDR 0.05)	
# of Up-regulated genes	171	179
# of Down-regulated genes	45	54
Total number of changed genes	216*	233

*These genes are reported in the table for differentially expressed genes (DEGs) and shown in the DEG heat maps.

Summary heat map showing all the differentially expressed genes (DEGs)

Here log gene expression levels were scaled and used to cluster genes and samples. Each row is a gene, each column is a sample. Gene names are not shown in this plot. Click image to download PDF version.

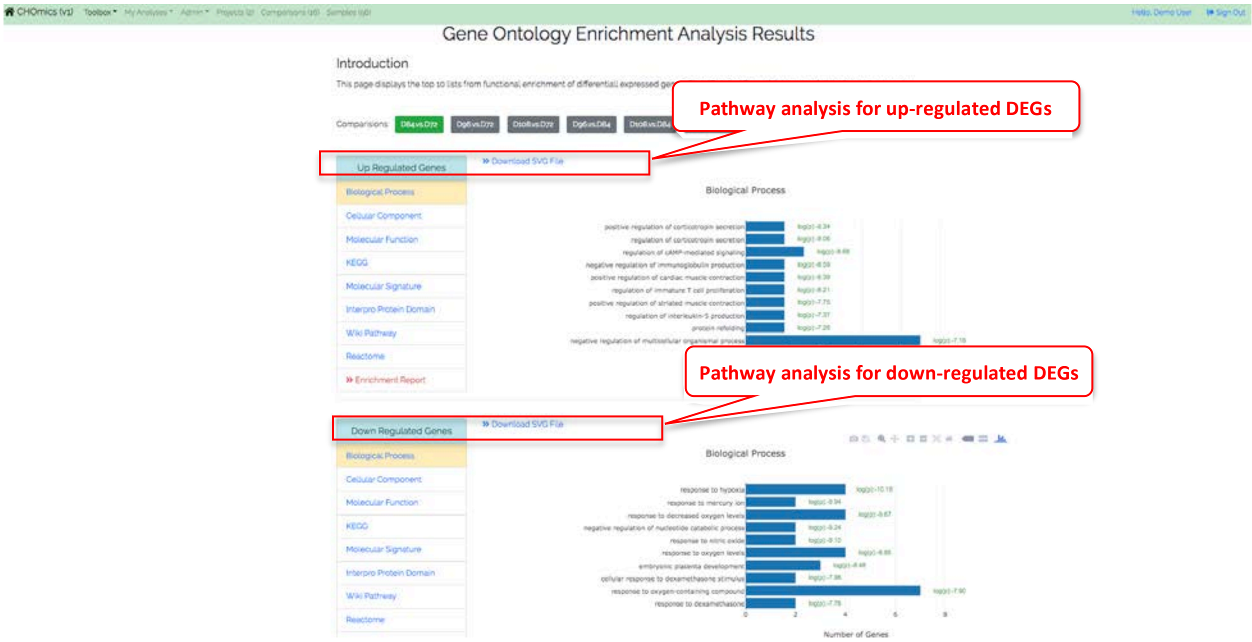
» View detailed heat map showing names for all DEGs (If there are more than 1000 DEGs, only the top 1000 genes are shown)

Sometimes, you may want to view gene-only clustering while leaving the sample order unchanged. You can view such heat maps here: Summary or Detailed Plots

DEGs summary table

DEGs Expression heatmap

In the report for GO enrichment analysis, barplots show the significance of enrichment of up- or down-regulated DEGs in different pathway databases, e.g, GO, KEGG, Wiki pathways, etc.



Gene Ontology Enrichment Analysis Results

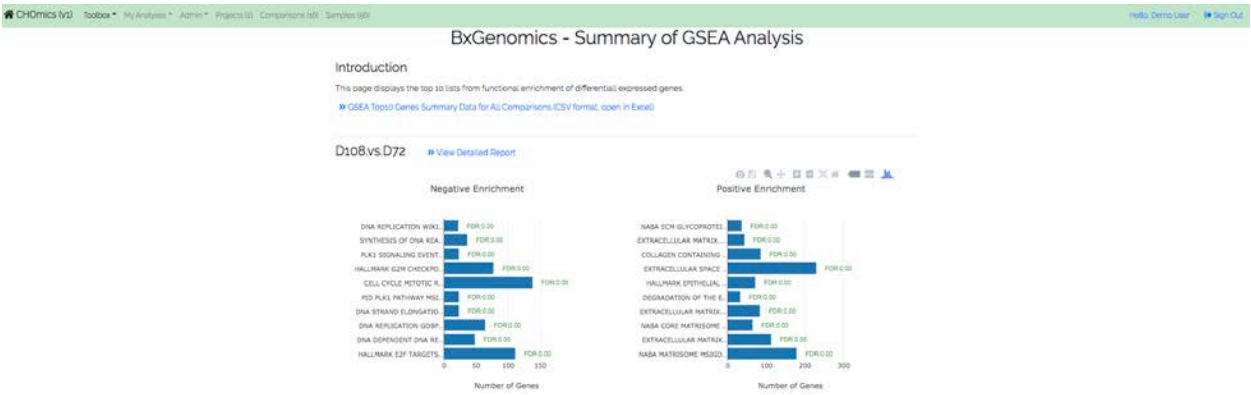
Introduction

This page displays the top 10 lists from functional enrichment of differentiall expressed ge

Pathway analysis for up-regulated DEGs

Pathway analysis for down-regulated DEGs

Similarly, in the report for GSEA analysis, enrichment results for up- and down- regulated DEGs in pathways from MigDB database are plotted with significance level (FDR), respectively.



BxGenomics - Summary of GSEA Analysis

Introduction

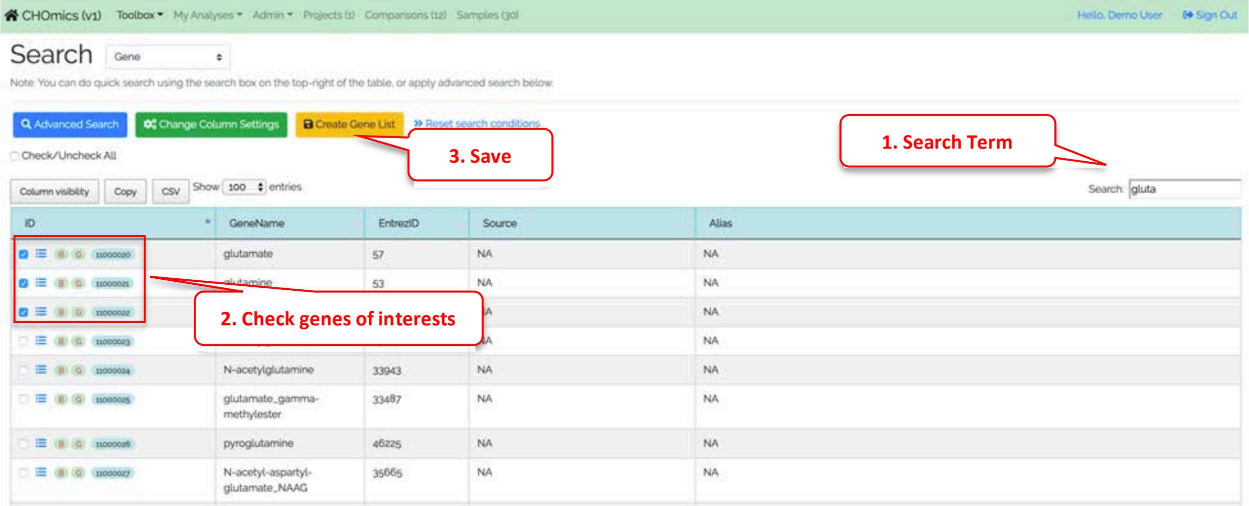
This page displays the top 10 lists from functional enrichment of differential expressed genes.

» GSEA Top10 Genes Summary Data for All Comparisons (CSV format, open in Excel)

D108.vs.D72 » View Detailed Report

3.3 Saved Genes and Comparisons

Customers can save selected genes or comparison for future use (e.g. multiple gene and multiple comparisons bubble plot). From gene search, check the genes you want to save, and click the yellow button "save selected genes



Search: Gene

Note: You can do quick search using the search box on the top-right of the table, or apply advanced search below.

1. Search Term

3. Save

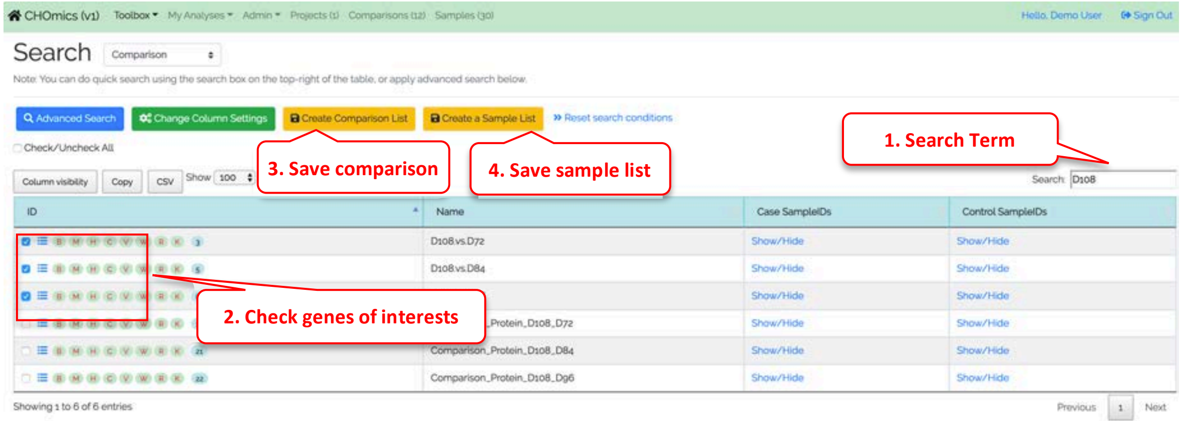
2. Check genes of interests

Show 100 entries

Search: gluta

ID	GeneName	EntrezID	Source	Alias
☑ 11000020	glutamate	57	NA	NA
☑ 11000021	glutamine	53	NA	NA
☑ 11000022				NA
☐ 11000023				NA
☐ 11000024	N-acetylglutamine	33943	NA	NA
☐ 11000025	glutamate_gamma-methylester	33487	NA	NA
☐ 11000026	pyroglutamine	46225	NA	NA
☐ 11000027	N-acetyl-aspartyl-glutamate_NAAG	35665	NA	NA

You can do the same with comparisons.



Search: Comparison

Note: You can do quick search using the search box on the top-right of the table, or apply advanced search below.

1. Search Term

3. Save comparison

4. Save sample list

2. Check genes of interests

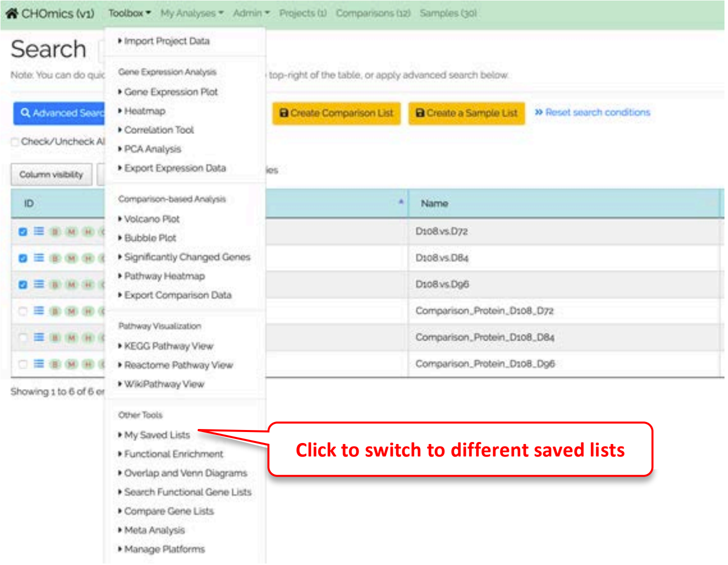
Show 100

Search: D108

ID	Name	Case SampleIDs	Control SampleIDs
☑ 3	D108.vs.D72	Show/Hide	Show/Hide
☑ 5	D108.vs.D84	Show/Hide	Show/Hide
☑		Show/Hide	Show/Hide
☐	_Protein_D108_D72	Show/Hide	Show/Hide
☐ 21	Comparison_Protein_D108_D84	Show/Hide	Show/Hide
☐ 22	Comparison_Protein_D108_D96	Show/Hide	Show/Hide

Showing 1 to 6 of 6 entries

To view selected genes or comparison, click “My Results” link on the left menu and then “Gene Lists”.

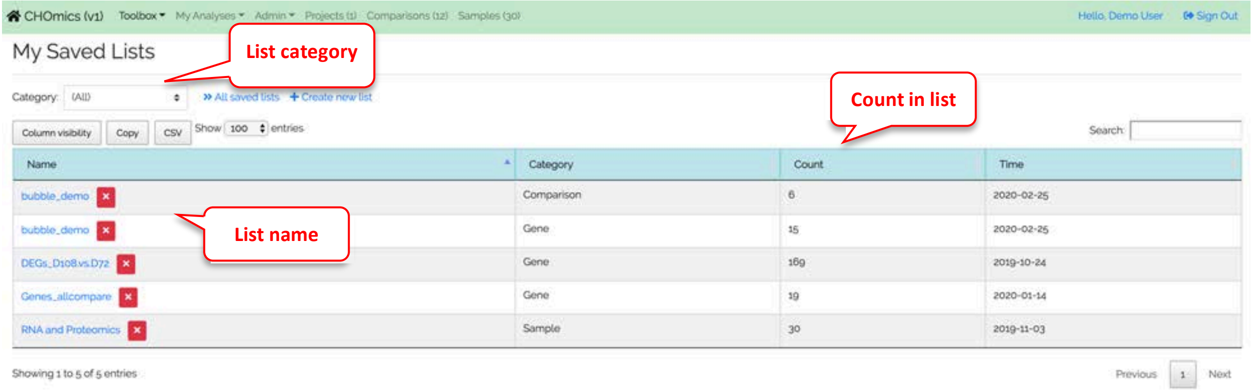


Note: You can do quic… top-right of the table, or apply advanced search below

ID	Name
☑	D108.vs.D72
☑	D108.vs.D84
☑	D108.vs.D96
□	Comparison_Protein_D108_D72
□	Comparison_Protein_D108_D84
□	Comparison_Protein_D108_D96

Showing 1 to 6 of 6 en…

Click to switch to different saved lists



My Saved Lists

List category

Category: (All)

Count in list

Show 100 entries

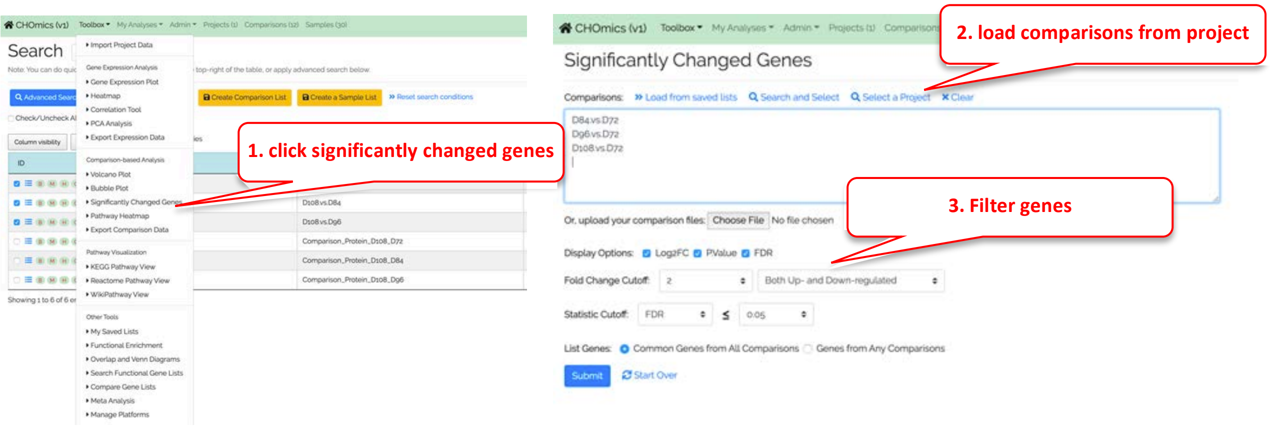
Search:

Name	Category	Count	Time
bubble_demo	Comparison	6	2020-02-25
bubble_demo	Gene	15	2020-02-25
DEGs_D108.vs.D72	Gene	169	2019-10-24
Genes_allcompare	Gene	19	2020-01-14
RNA and Proteomics	Sample	30	2019-11-03

List name

Showing 1 to 5 of 5 entries

One additional way to select and save genes or comparisons is from the ‘significant changed genes’ in toolbox. Using the dynamic filters to choose the comparisons or genes you are interested in, and you can use the table at the bottom to save comparisons or genes.



1. click significantly changed genes

2. load comparisons from project

3. Filter genes

Significantly Changed Genes

Comparisons:

D84.vs.D72
D96.vs.D72
D108.vs.D72

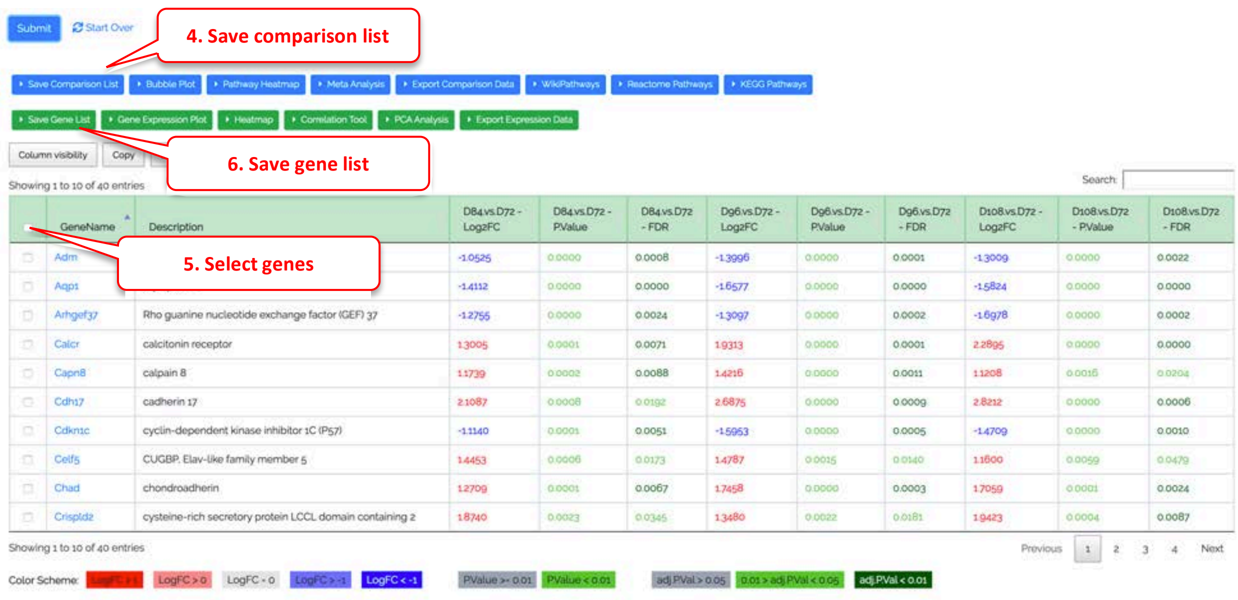
Or, upload your comparison files: No file chosen

Display Options: ☑ Log2FC ☑ PValue ☑ FDR

Fold Change Cutoff: 2 Both Up- and Down-regulated

Statistic Cutoff: FDR ≤ 0.05

List Genes: ● Common Genes from All Comparisons ○ Genes from Any Comparisons



Showing 1 to 10 of 40 entries

	GeneName	Description	D84.vs.D72 - Log2FC	D84.vs.D72 - P.Value	D84.vs.D72 - FDR	D96.vs.D72 - Log2FC	D96.vs.D72 - P.Value	D96.vs.D72 - FDR	D108.vs.D72 - Log2FC	D108.vs.D72 - P.Value	D108.vs.D72 - FDR
□	Adm		-1.0525	0.0000	0.0008	-1.3996	0.0000	0.0001	-1.3009	0.0000	0.0022
□	Aqp1		-1.4112	0.0000	0.0000	-1.6577	0.0000	0.0000	-1.5824	0.0000	0.0000
□	Arhgef37	Rho guanine nucleotide exchange factor (GEF) 37	-1.2755	0.0000	0.0024	-1.3097	0.0000	0.0002	-1.6978	0.0000	0.0002
□	Calcr	calcitonin receptor	1.3005	0.0001	0.0071	1.9313	0.0000	0.0001	2.2895	0.0000	0.0000
□	Capn8	calpain 8	1.1739	0.0002	0.0088	1.4216	0.0000	0.0011	1.1208	0.0016	0.0204
□	Cdh17	cadherin 17	2.1087	0.0008	0.0192	2.6875	0.0000	0.0009	2.8212	0.0000	0.0006
□	Cdkn1c	cyclin-dependent kinase inhibitor 1C (P57)	-1.1140	0.0001	0.0051	-1.5953	0.0000	0.0005	-1.4709	0.0000	0.0010
□	Celf5	CUGBP, Elav-like family member 5	1.4453	0.0006	0.0173	1.4787	0.0015	0.0140	1.1600	0.0059	0.0479
□	Chad	chondroadherin	1.2709	0.0001	0.0067	1.7458	0.0000	0.0003	1.7059	0.0001	0.0024
□	Crispld2	cysteine-rich secretory protein LCCL domain containing 2	1.8740	0.0023	0.0345	1.3480	0.0022	0.0181	1.9423	0.0004	0.0087

Showing 1 to 10 of 40 entries

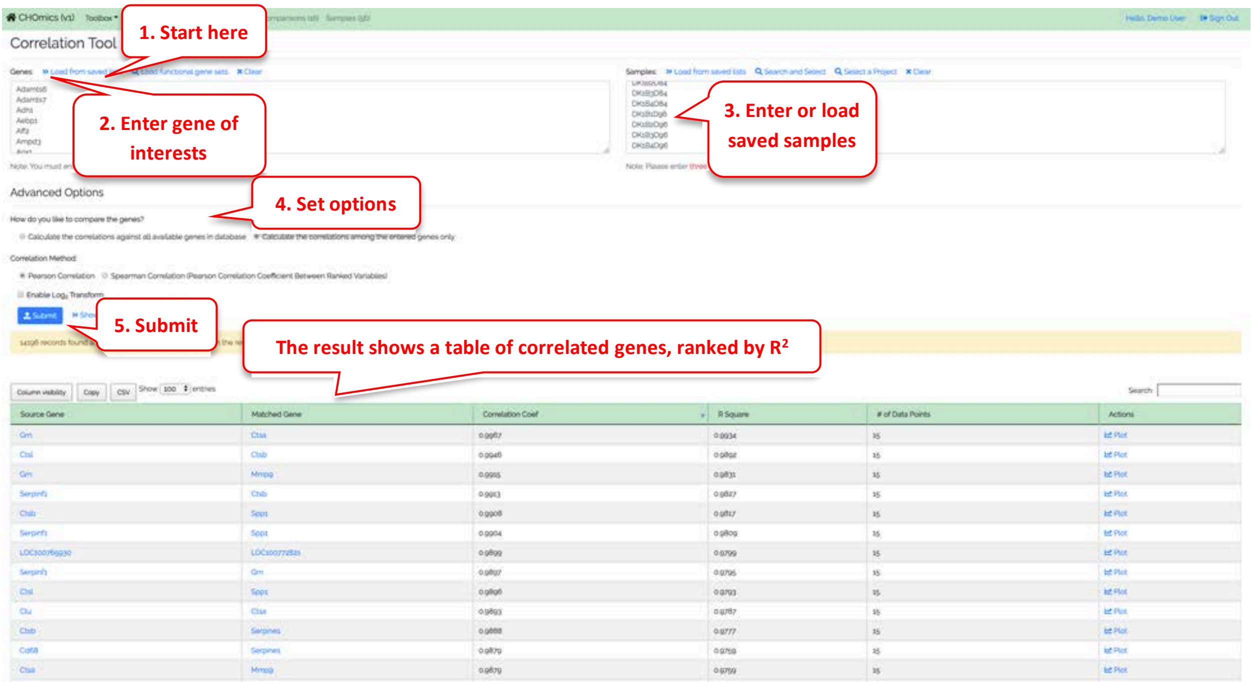
Color Scheme: LogFC > 1 LogFC > 0 LogFC = 0 LogFC > -1 LogFC < -1 PValue >= 0.01 PValue < 0.01 adj.PVal > 0.05 0.01 > adj.PVal < 0.05 adj.PVal < 0.01

3.4 Advanced Analysis

Besides the above analyses, the CHOmics also provides several advanced tools.

3.3.1 Correlation Tools

Once the user has identified a gene of interest, the user can use correlation tools to find other genes that share similar (or opposite) profiles in terms of gene expression or fold change. First, enter the gene of interest, and samples to be used for correlation. In the example below, we entered a saved gene list, and 15 samples.



Correlation Tool

Genes: Adamts6, Adamts7, Adhs, Aebp1, Aff2, Ampd3

Samples: DK1B3D84, DK1B4D84, DK1B1D96, DK1B2D96, DK1B3D96, DK1B4D96

Advanced Options

How do you like to compare the genes?

○ Calculate the correlations against all available genes in database ● Calculate the correlations among the entered genes only

Correlation Method:

● Pearson Correlation ○ Spearman Correlation (Pearson Correlation Coefficient Between Ranked Variables)

□ Enable Log_2 Transform

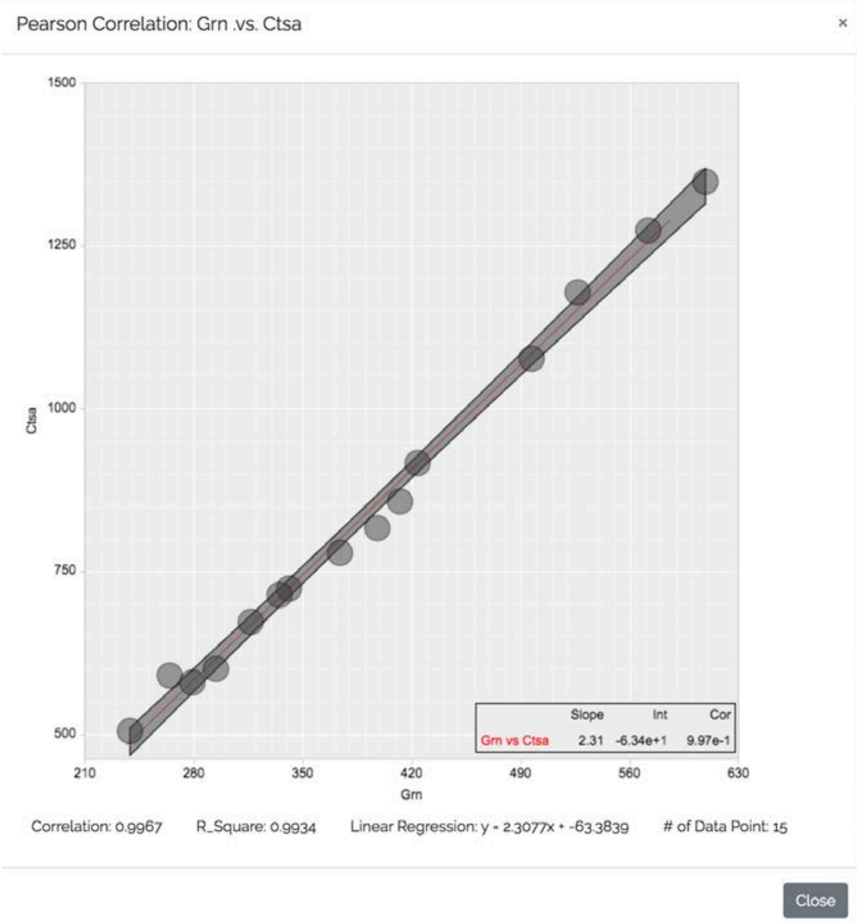
14196 records found

The result shows a table of correlated genes, ranked by R^2

Show 100 entries

Source Gene	Matched Gene	Correlation Coef	R Square	# of Data Points	Actions
Grn	Ctsa	0.9967	0.9934	15	Plot
Ctsl	Ctsb	0.9946	0.9892	15	Plot
Grn	Mmp9	0.9915	0.9831	15	Plot
Serpinf1	Ctsb	0.9913	0.9827	15	Plot
Ctsb	Spp1	0.9908	0.9817	15	Plot
Serpinf1	Spp1	0.9904	0.9809	15	Plot
LOC100769930	LOC100772821	0.9899	0.9799	15	Plot
Serpinf1	Grn	0.9897	0.9795	15	Plot
Ctsl	Spp1	0.9896	0.9793	15	Plot
Clu	Ctsa	0.9893	0.9787	15	Plot
Ctsb	Serpine1	0.9888	0.9777	15	Plot
Cd68	Serpine1	0.9879	0.9759	15	Plot
Ctsa	Mmp9	0.9879	0.9759	15	Plot

Click the plot icon will show scatter plot of the target and the correlated gene (e.g, gene Grn vs. Ctsa).

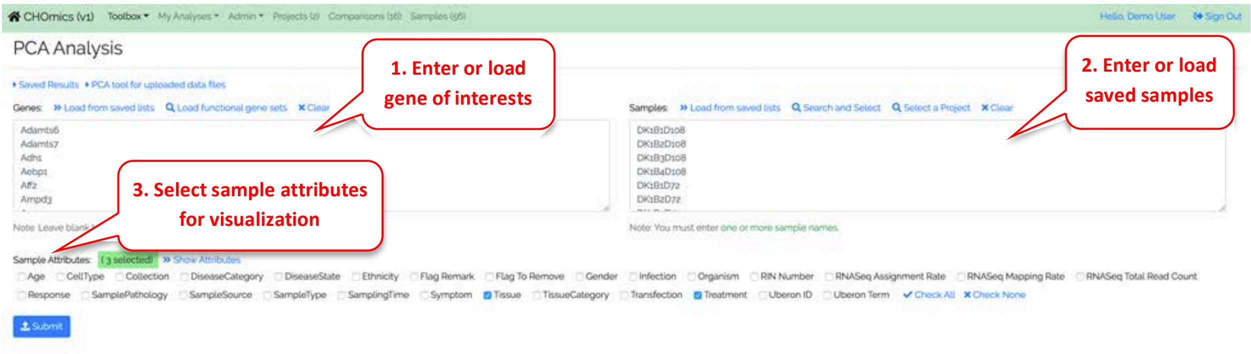


Pearson Correlation: Grn .vs. Ctsa ×

Correlation: 0.9967 R_Square: 0.9934 Linear Regression: y = 2.3077x + -63.3839 # of Data Point: 15

3.3.2 PCA Analysis

You can select a set of samples and genes and use PCA plot to visualize the sample relationships on the target gene set.



CHOmics (v1) Toolbox ▾ My Analyses ▾ Admin ▾ Projects (2) Comparisons (16) Samples (56) Hello, Demo User Sign Out

PCA Analysis

▸ Saved Results ▸ PCA tool for uploaded data files

Genes:

» Load from saved lists Load functional gene sets ✕ Clear

Adamts6
Adamts7
Adh1
Aebp1
Aff2
Ampd3

Note: Leave blank

Samples:

» Load from saved lists Search and Select Select a Project ✕ Clear

DK1B1D108
DK1B2D108
DK1B3D108
DK1B4D108
DK1B1D72
DK1B2D72

Note: You must enter one or more sample names.

Sample Attributes: (3 selected)

□ Age □ CellType □ Collection □ DiseaseCategory □ DiseaseState □ Ethnicity □ Flag Remark □ Flag To Remove □ Gender □ Infection □ Organism □ RIN Number □ RNASeq Assignment Rate □ RNASeq Mapping Rate □ RNASeq Total Read Count □ Response □ SamplePathology □ SampleSource □ SampleType □ SamplingTime □ Symptom ☑ Tissue □ TissueCategory □ Transfection ☑ Treatment □ Uberon ID □ Uberon Term

1. Enter or load gene of interests

2. Enter or load saved samples

3. Select sample attributes for visualization

The system will use FactorMineR package to run PCA analysis and display the results. Several PCA metrics are plotted for interactive visualization:

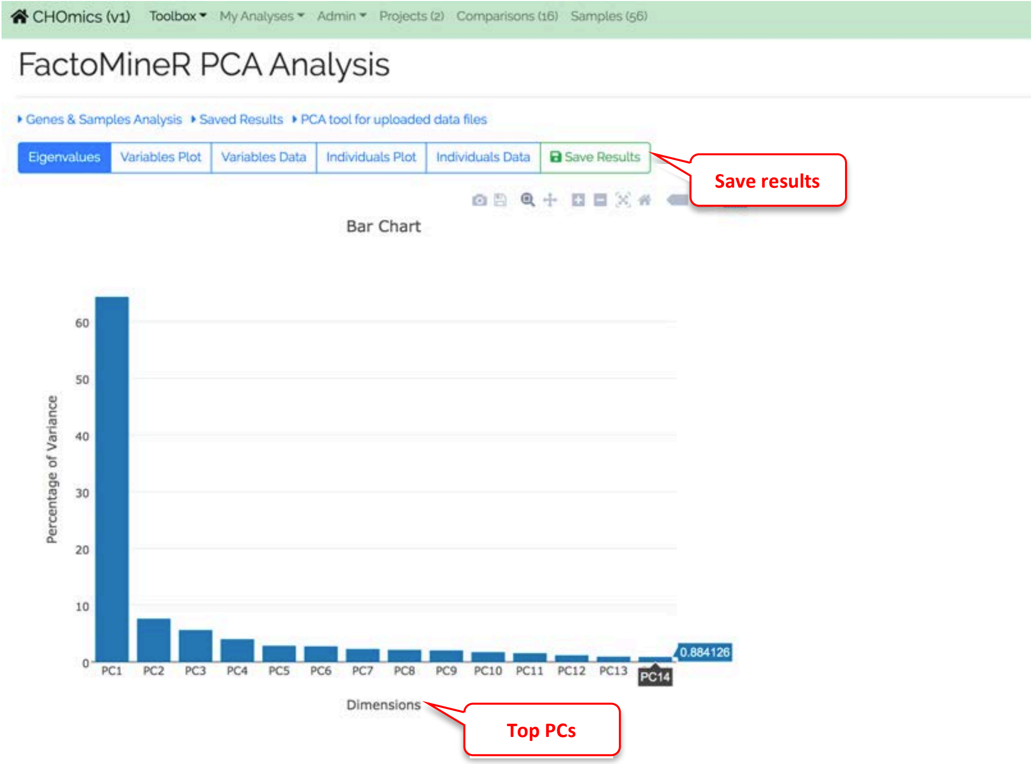
Eigenvalues plot the percentage of variance explained by top PCs.

Variables Plot shows the weights of top contributing genes in each PCs.

Variable Data summarizes the weights of each gene in each individual PC.

Individuals Plot shows the relationship of samples on the spanned space by different PCs.

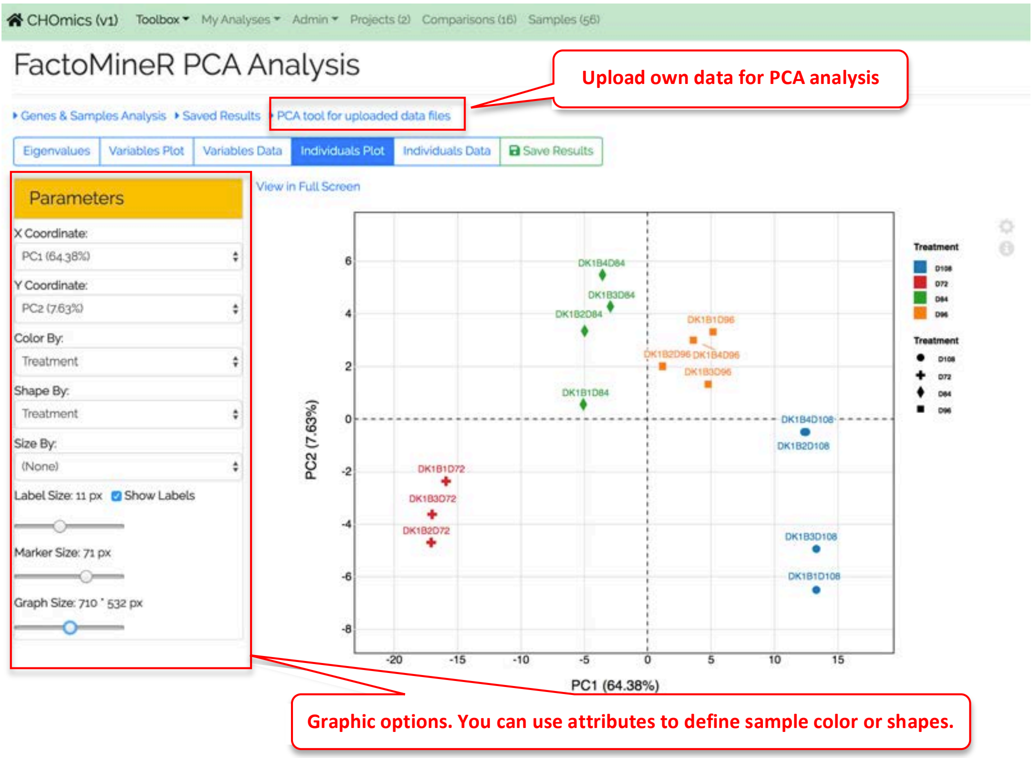
Individual Data summarizes the score vector of each sample in each individual PC.



FactoMineR PCA Analysis

Save results

Top PCs



FactoMineR PCA Analysis

Upload own data for PCA analysis

X Coordinate: PC1 (64.38%)

Y Coordinate: PC2 (7.63%)

Color By: Treatment

Shape By: Treatment

Size By: (None)

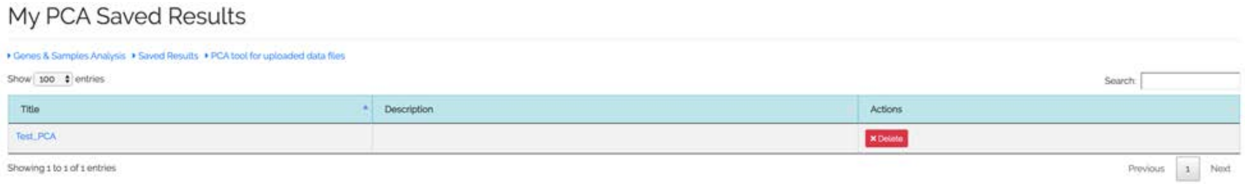
Label Size: 11 px ☑ Show Labels

Marker Size: 71 px

Graph Size: 710 * 532 px

Graphic options. You can use attributes to define sample color or shapes.

The PCA results can be saved. Users can load it in the future.



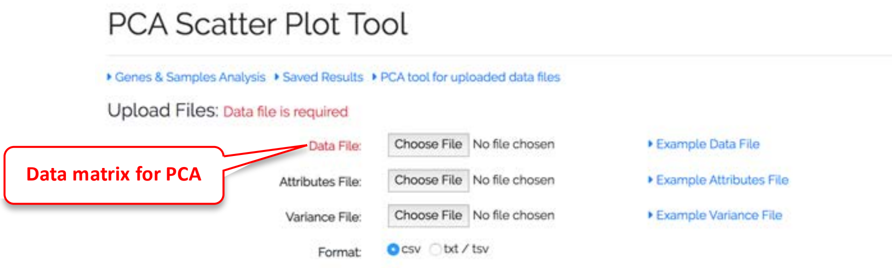
My PCA Saved Results

Show 100 entries

Title	Description	Actions
Test_PCA		✖ Delete

Showing 1 to 1 of 1 entries

Users can upload their own data matrix or pre-calculated data for PCA analysis and visualization.



PCA Scatter Plot Tool

Upload Files: Data file is required

Data File: No file chosen

Attributes File: No file chosen

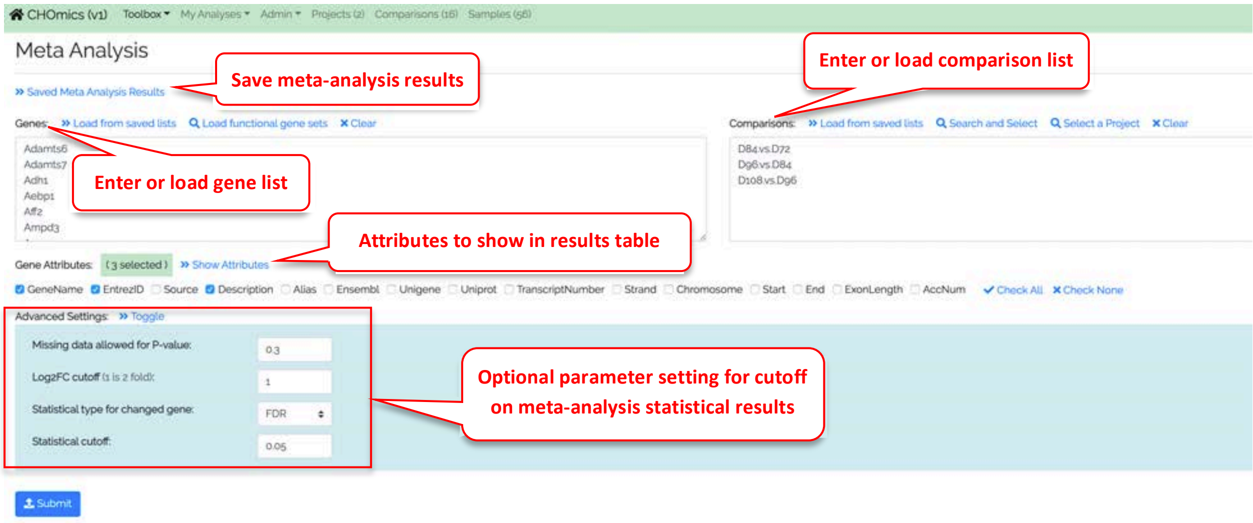
Variance File: No file chosen

Format: ● csv ○ txt / tsv

Data matrix for PCA

3.3.3 Meta-Analysis

Meta-Analysis can be used to identify genes that are changed consistently across multiple projects. It is listed as one functional module in toolbox panel. In the example below, we are looking for the most significant DEGs in three comparisons.



Meta Analysis

» Saved Meta Analysis Results

Save meta-analysis results

Genes: » Load from saved lists · Load functional gene sets · Clear

Adamts6
Adamts7
Adh1
Aebp1
Aff2
Ampd3

Enter or load gene list

Comparisons: » Load from saved lists · Search and Select · Select a Project · Clear

D84.vs.D72
D96.vs.D84
D108.vs.D96

Enter or load comparison list

Gene Attributes: (3 selected) » Show Attributes

Attributes to show in results table

☑ GeneName ☑ EntrezID □ Source ☑ Description □ Alias □ Ensembl □ Unigene □ Uniprot □ TranscriptNumber □ Strand □ Chromosome □ Start □ End □ ExonLength □ AccNum ✓ Check All ✕ Check None

Advanced Settings: » Toggle

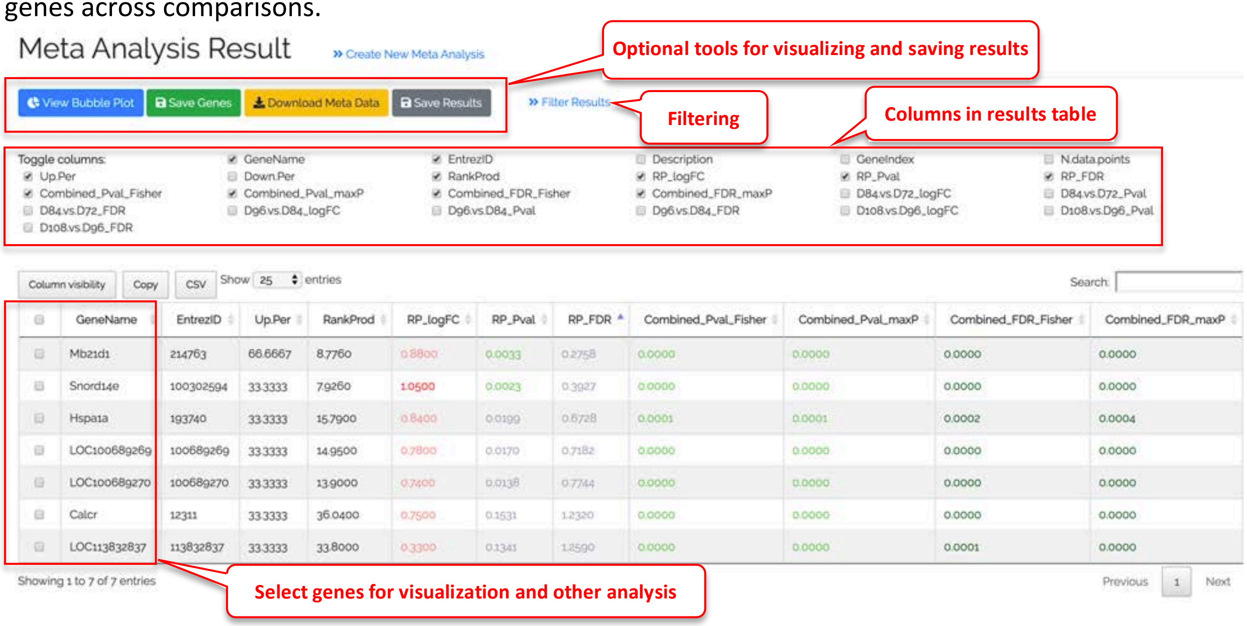
Missing data allowed for P-value:	0.3
Log2FC cutoff (1 is 2 fold):	1
Statistical type for changed gene:	FDR
Statistical cutoff:	0.05

Optional parameter setting for cutoff on meta-analysis statistical results

The meta-analysis pipeline will compute three types of results:

1) Maximum p-value (maxP). This method targets on DEGs have small *p*-values in "all" comparisons. We recommend using maxP if you are looking for DEGs that are common among several studies.
2) Fisher's p-value. The Fisher's method sums up the log-transformed p-values obtained from individual studies. This p-value combination method is useful if you want to identify DEGs in any of the comparisons.
3) We also applied simple counting method to report the frequency a gene is classified as up or down-regulated DEG from all the comparisons. The default DEG cutoff is two-fold change and FDR<0.05. but user can change the cutoff.

In most cases, combing maxP (smaller values are more significant) and the counting method (e.g. up-regulated in 50% of studies) will give the most biological relevant results for consistently regulated genes across comparisons.



Meta Analysis Result

» Create New Meta Analysis

Optional tools for visualizing and saving results

» Filter Results

Filtering

Columns in results table

Toggle columns:

☑ Up.Per
☑ Combined_Pval_Fisher
□ D84.vs.D72_FDR
□ D108.vs.D96_FDR
☑ GeneName
□ Down.Per
☑ Combined_Pval_maxP
□ D96.vs.D84_logFC
☑ EntrezID
☑ RankProd
☑ Combined_FDR_Fisher
□ D96.vs.D84_Pval
□ Description
☑ RP_logFC
☑ Combined_FDR_maxP
□ D96.vs.D84_FDR
□ GeneIndex
☑ RP_Pval
□ D84.vs.D72_logFC
□ D108.vs.D96_logFC
□ N.data.points
☑ RP_FDR
□ D84.vs.D72_Pval
□ D108.vs.D96_Pval

Show 25 entries

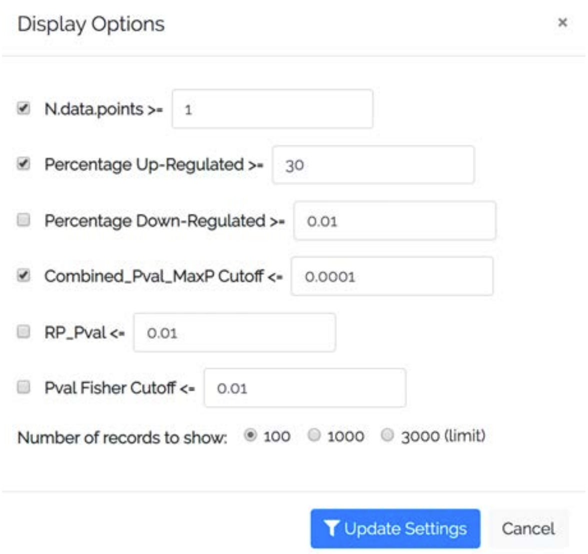
Search:

□	GeneName	EntrezID	Up.Per	RankProd	RP_logFC	RP_Pval	RP_FDR	Combined_Pval_Fisher	Combined_Pval_maxP	Combined_FDR_Fisher	Combined_FDR_maxP
□	Mb21d1	214763	66.6667	8.7760	0.8800	0.0033	0.2758	0.0000	0.0000	0.0000	0.0000
□	Snord14e	100302594	33.3333	7.9260	1.0500	0.0023	0.3927	0.0000	0.0000	0.0000	0.0000
□	Hspa1a	193740	33.3333	15.7900	0.8400	0.0199	0.6728	0.0001	0.0001	0.0002	0.0004
□	LOC100689269	100689269	33.3333	14.9500	0.7800	0.0170	0.7182	0.0000	0.0000	0.0000	0.0000
□	LOC100689270	100689270	33.3333	13.9000	0.7400	0.0138	0.7744	0.0000	0.0000	0.0000	0.0000
□	Calcr	12311	33.3333	36.0400	0.7500	0.1531	1.2320	0.0000	0.0000	0.0000	0.0000
□	LOC113832837	113832837	33.3333	33.8000	0.3300	0.1341	1.2590	0.0000	0.0000	0.0001	0.0000

Showing 1 to 7 of 7 entries

Select genes for visualization and other analysis

In the above example, we used a relatively loose filtering criterion (N.data.points>1, and up-regulation in percentage>30% of studies, and Combined_Pval_MaxP <=0.0001) because only small number of genes pass the stringent default criterion.



Display Options ×

☑ N.data.points >= 1

☑ Percentage Up-Regulated >= 30

□ Percentage Down-Regulated >= 0.01

☑ Combined_Pval_MaxP Cutoff <= 0.0001

□ RP_Pval <= 0.01

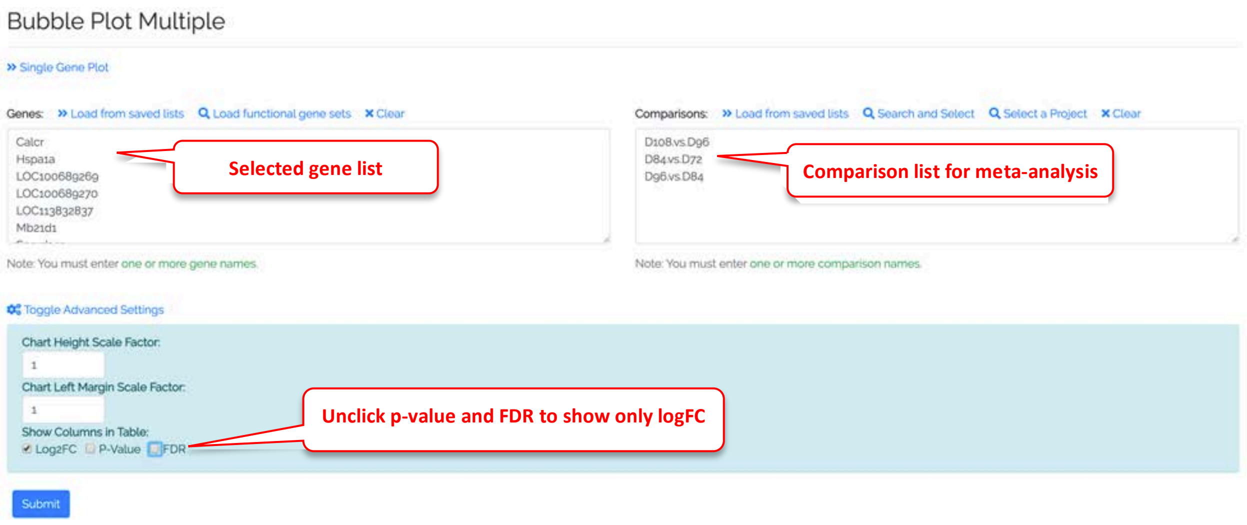
□ Pval Fisher Cutoff <= 0.01

Number of records to show: ● 100 ○ 1000 ○ 3000 (limit)

The data table shows the genes that pass the filters. We can sort the table by maxP value. A different filter can be applied to get down-regulated genes.

The results can be saved for future access. There are also links to several other tools. The download meta data link will save a CSV file that contain results from all genes.

Next, we will choose all the genes that pass filter by checking the box for all listed genes, and use bubble plot to visualize the results.



Bubble Plot Multiple

Genes:

Calcr
Hspa1a
LOC100689269
LOC100689270
LOC113832837
Mb21d1

Note: You must enter one or more gene names.

Comparisons:

D108.vs.D96
D84.vs.D72
D96.vs.D84

Note: You must enter one or more comparison names.

Chart Height Scale Factor: 1

Chart Left Margin Scale Factor: 1

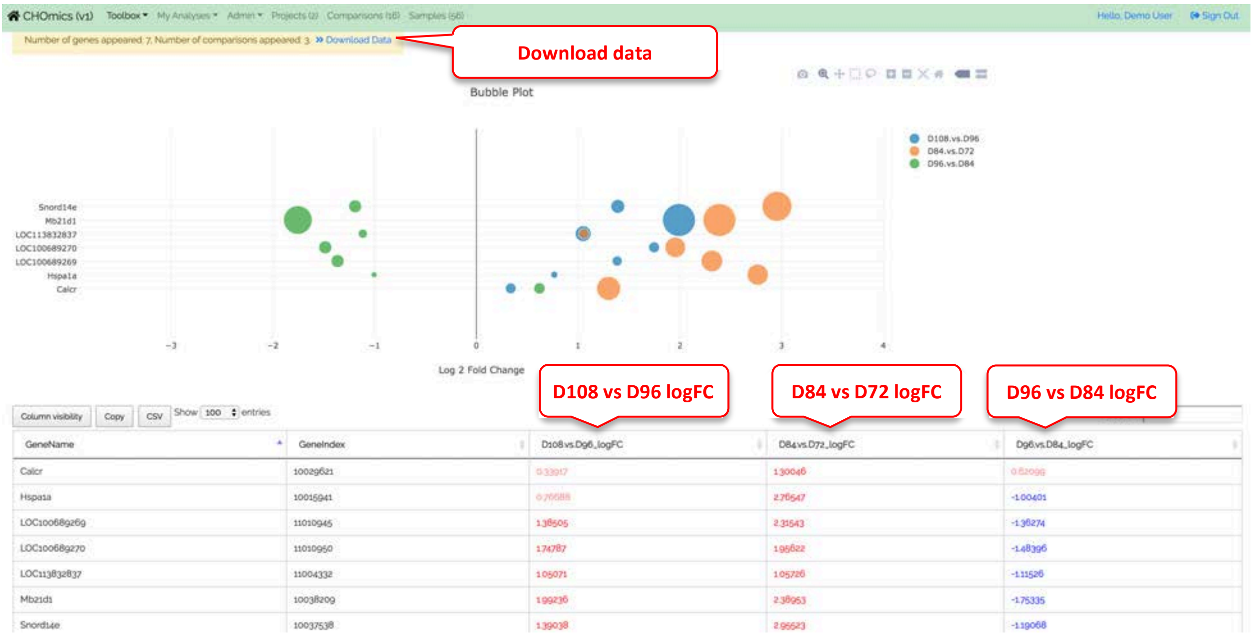
Show Columns in Table: ☑ Log2FC □ P-Value □ FDR

Selected gene list

Comparison list for meta-analysis

Unclick p-value and FDR to show only logFC

The resulting bubble plot will show all three comparisons for each gene.



Number of genes appeared: 7, Number of comparisons appeared: 3. » Download Data

Download data

D108 vs D96 logFC

D84 vs D72 logFC

D96 vs D84 logFC

GeneName	GeneIndex	D108.vs.D96_logFC	D84.vs.D72_logFC	D96.vs.D84_logFC
Calcr	10029621	0.33917	1.30046	0.82099
Hspa1a	10015941	0.70688	2.76547	-1.00401
LOC100689269	11010945	1.38505	2.31543	-1.38274
LOC100689270	11010950	1.74787	1.95822	-1.48396
LOC113832837	11004332	1.05071	1.05726	-1.11526
Mb21d1	10038209	1.99236	2.38953	-1.75335
Snord14e	10037538	1.39038	2.95523	-1.19068

The data table below the bubble plot can also be used for filtering. Remember in the advanced settings, we choose to display logFC only, this makes it easier to look for genes that are reverted in different time points. The logFC values are colored coded (red, increase, blue, decrease), therefore we can see that most of genes show upregulation in D84, and then downregulation in D96 and then upregulation in D108.

You can also redo the plot, check all columns to include p-value and FDR in the table, and export the results to excel file.

The workflow above uses up-regulated genes as example. You can get down-regulated genes from the filter step in meta-analysis result page.

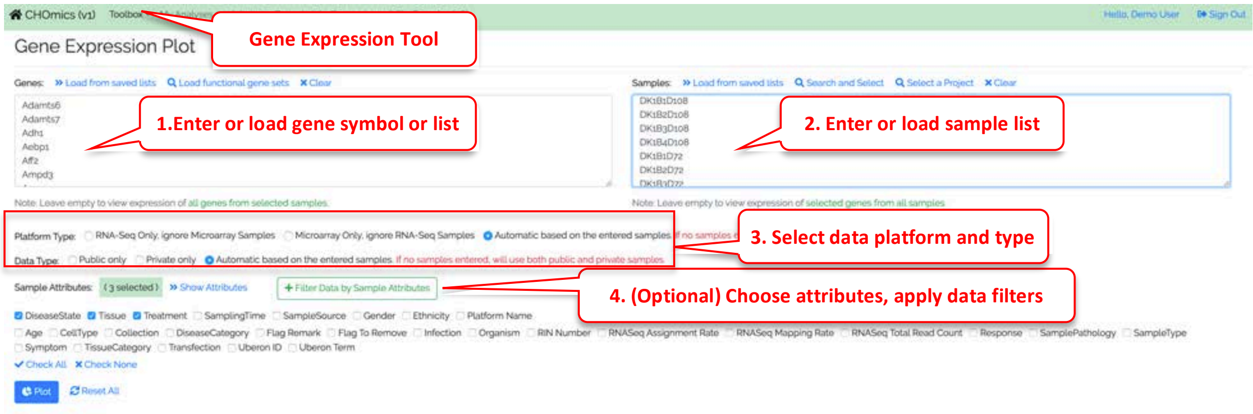
4 Visualization

4.1 Visualize Gene Expression

CHOmics provides tool to easily visualize gene expression level across multiple genes, samples and omics. For each gene, you can view its expression levels across multiple samples.

4.1.1 View Gene Expression from multiple samples

Choose the Gene Expression tool from Toolbox -> Gene Expression Plot from top menu, and enter the official symbol of genes or load gene list from saved lists. Alternatively, in the gene details page, click View Gene Expression link.



Gene Expression Tool

Gene Expression Plot

1.Enter or load gene symbol or list

Genes:
Adamts6
Adamts7
Adh1
Aebp1
Aff2
Ampd3

Note: Leave empty to view expression of all genes from selected samples.

2. Enter or load sample list

Samples:
DK1B1D108
DK1B2D108
DK1B3D108
DK1B4D108
DK1B1D72
DK1B2D72
DK1B3D72

Note: Leave empty to view expression of selected genes from all samples

3. Select data platform and type

Platform Type: ○ RNA-Seq Only, ignore Microarray Samples ○ Microarray Only, ignore RNA-Seq Samples ● Automatic based on the entered samples. If no samples e

Data Type: ○ Public only ○ Private only ● Automatic based on the entered samples. If no samples entered, will use both public and private samples.

4. (Optional) Choose attributes, apply data filters

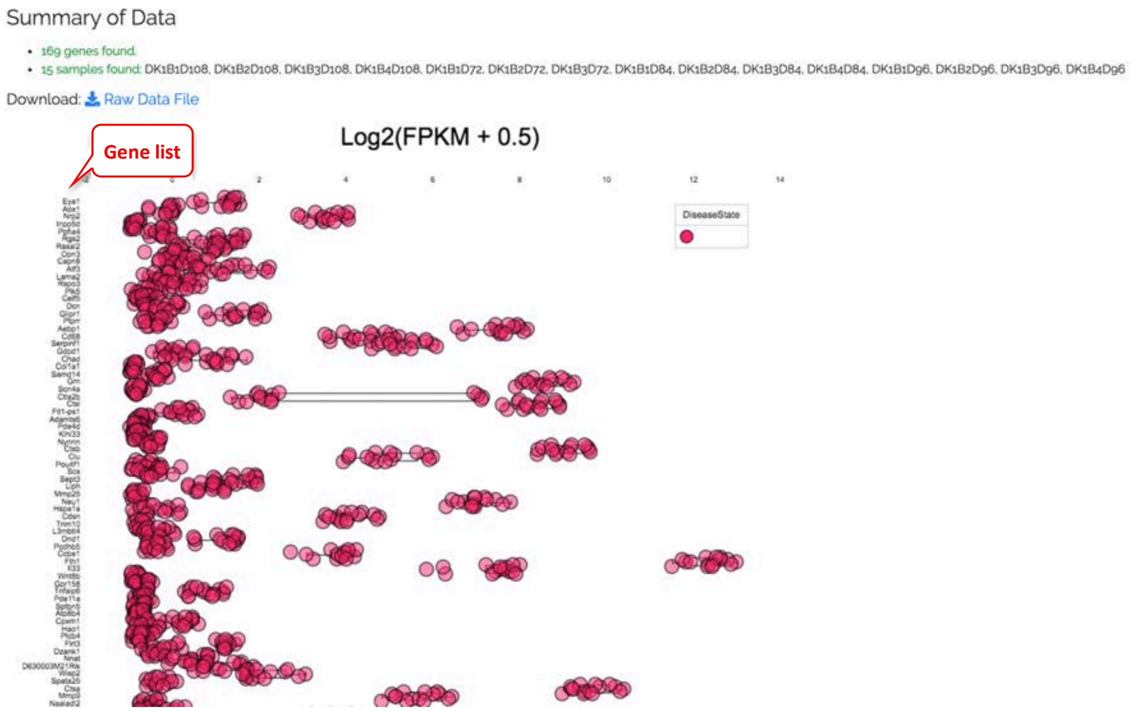
Sample Attributes: (3 selected)

☑ DiseaseState ☑ Tissue ☑ Treatment □ SamplingTime □ SampleSource □ Gender □ Ethnicity □ Platform Name
□ Age □ CellType □ Collection □ DiseaseCategory □ Flag Remark □ Flag To Remove □ Infection □ Organism □ RIN Number □ RNASeq Assignment Rate □ RNASeq Mapping Rate □ RNASeq Total Read Count □ Response □ SamplePathology □ SampleType
□ Symptom □ TissueCategory □ Transfection □ Uberon ID □ Uberon Term

As an optional step, you can choose what sample attributes to pass to the plot, and use data filter to choose only a subset of data points.

The Data Filter can be very useful if there are too many data points, and you want to focus on a few diseases or tissue types.

The screenshots below show default boxplot showing all samples by different time points (i.e, treatment).



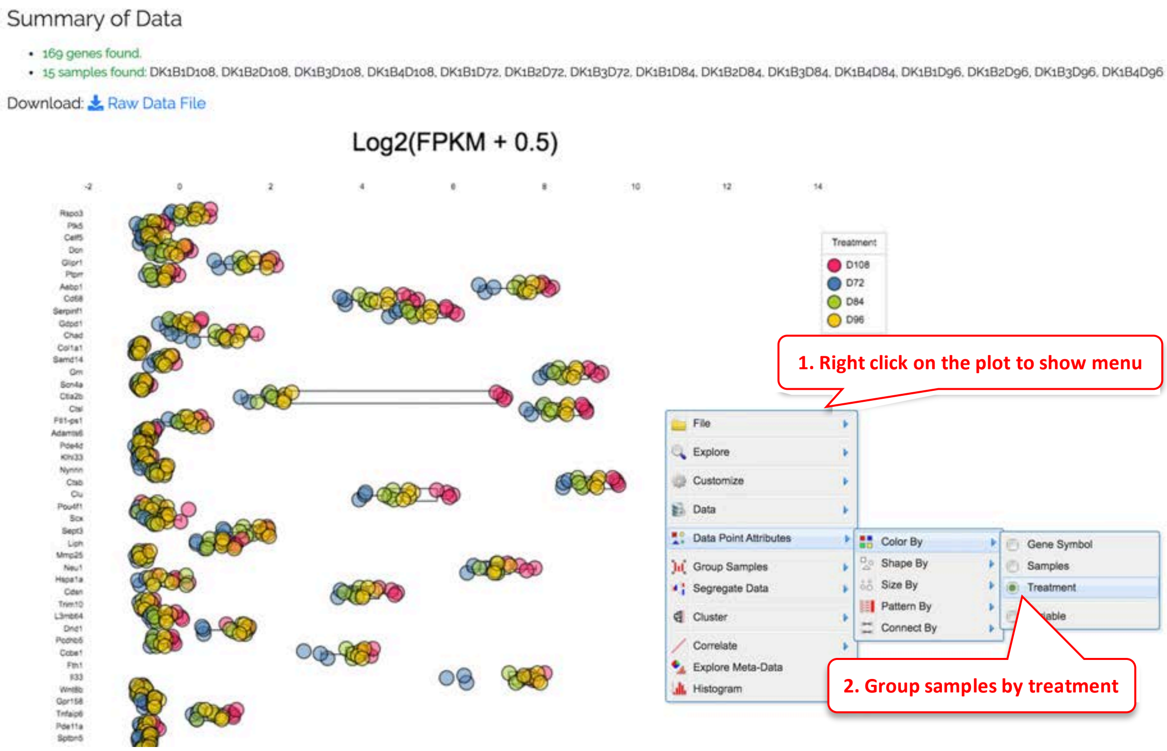
Summary of Data

- 169 genes found.
- 15 samples found: DK1B1D108, DK1B2D108, DK1B3D108, DK1B4D108, DK1B1D72, DK1B2D72, DK1B3D72, DK1B1D84, DK1B2D84, DK1B3D84, DK1B4D84, DK1B1D96, DK1B2D96, DK1B3D96, DK1B4D96

Download: Raw Data File

Customize Gene Expression Plot

The boxplot is created using CanvasXpress (https://canvasxpress.org)plug-in, and sample grouping and coloring can be customized by the user. In the example below, we show how data points are colored.



Summary of Data

- 169 genes found.
- 15 samples found: DK1B1D108, DK1B2D108, DK1B3D108, DK1B4D108, DK1B1D72, DK1B2D72, DK1B3D72, DK1B1D84, DK1B2D84, DK1B3D84, DK1B4D84, DK1B1D96, DK1B2D96, DK1B3D96, DK1B4D96

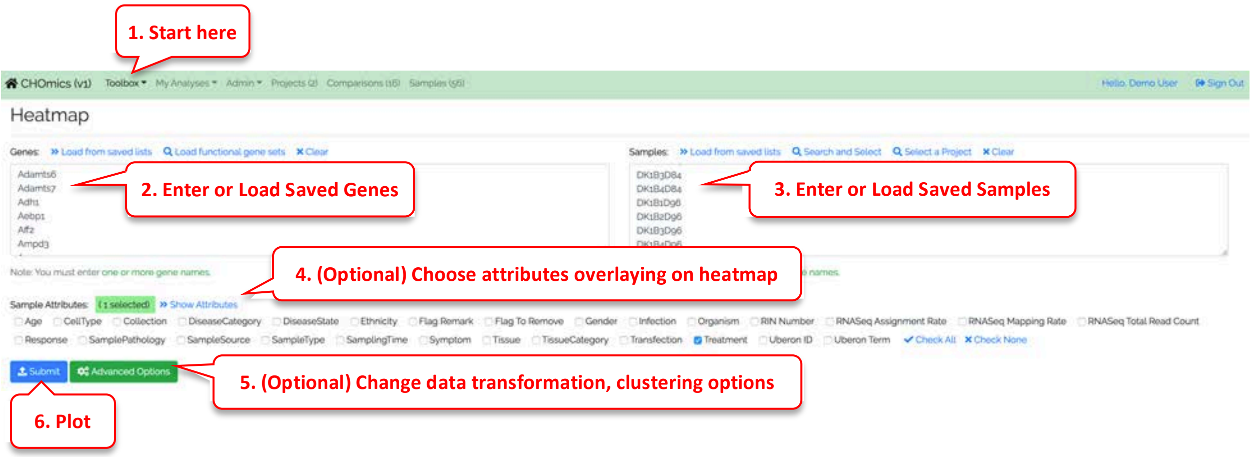
Download: Raw Data File

1. Right click on the plot to show menu

2. Group samples by treatment

4.1.2 View Gene Expression in Heatmap

Heatmap can be useful to visualize gene profiles from multiple samples. It can also provide information about how genes and samples cluster.



1. Start here

Heatmap

Genes:

Adamts6
Adamts7
Adh1
Aebp1
Aff2
Ampd3

2. Enter or Load Saved Genes

Samples:

DK1B3D84
DK1B4D84
DK1B1D98
DK1B2D96
DK1B3D96
DK1B4D96

3. Enter or Load Saved Samples

Note: You must enter one or more gene names.

4. (Optional) Choose attributes overlaying on heatmap

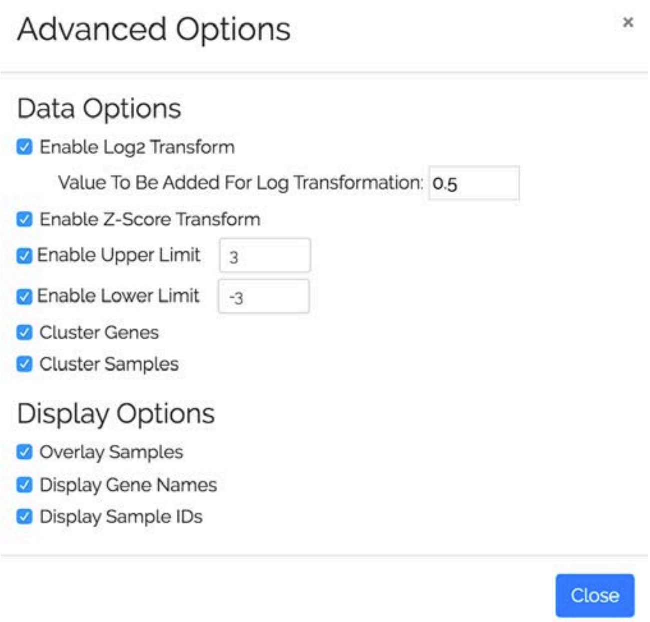
Sample Attributes: (1 selected)

□ Age □ CellType □ Collection □ DiseaseCategory □ DiseaseState □ Ethnicity □ Flag Remark □ Flag To Remove □ Gender □ Infection □ Organism □ RIN Number □ RNASeq Assignment Rate □ RNASeq Mapping Rate □ RNASeq Total Read Count
□ Response □ SamplePathology □ SampleSource □ SampleType □ SamplingTime □ Symptom □ Tissue □ TissueCategory □ Transfection ☑ Treatment □ Uberon ID □ Uberon Term

5. (Optional) Change data transformation, clustering options

6. Plot

You can enter genes and samples in the box, or load pre-saved genes and samples quickly from your collection. Be default, we will log2 transform the gene expression data, perform scaling of the data across samples for each gene, and limit the scaled value to -3 to 3 before displaying the data in heatmap. This works well in most situations. However, advanced users can change the options. For example, if you want to keep the order of samples as you entered, just uncheck "Cluster Samples".



Advanced Options

Data Options

☑ Enable Log2 Transform

Value To Be Added For Log Transformation: 0.5

☑ Enable Z-Score Transform

☑ Enable Upper Limit 3

☑ Enable Lower Limit -3

☑ Cluster Genes

☑ Cluster Samples

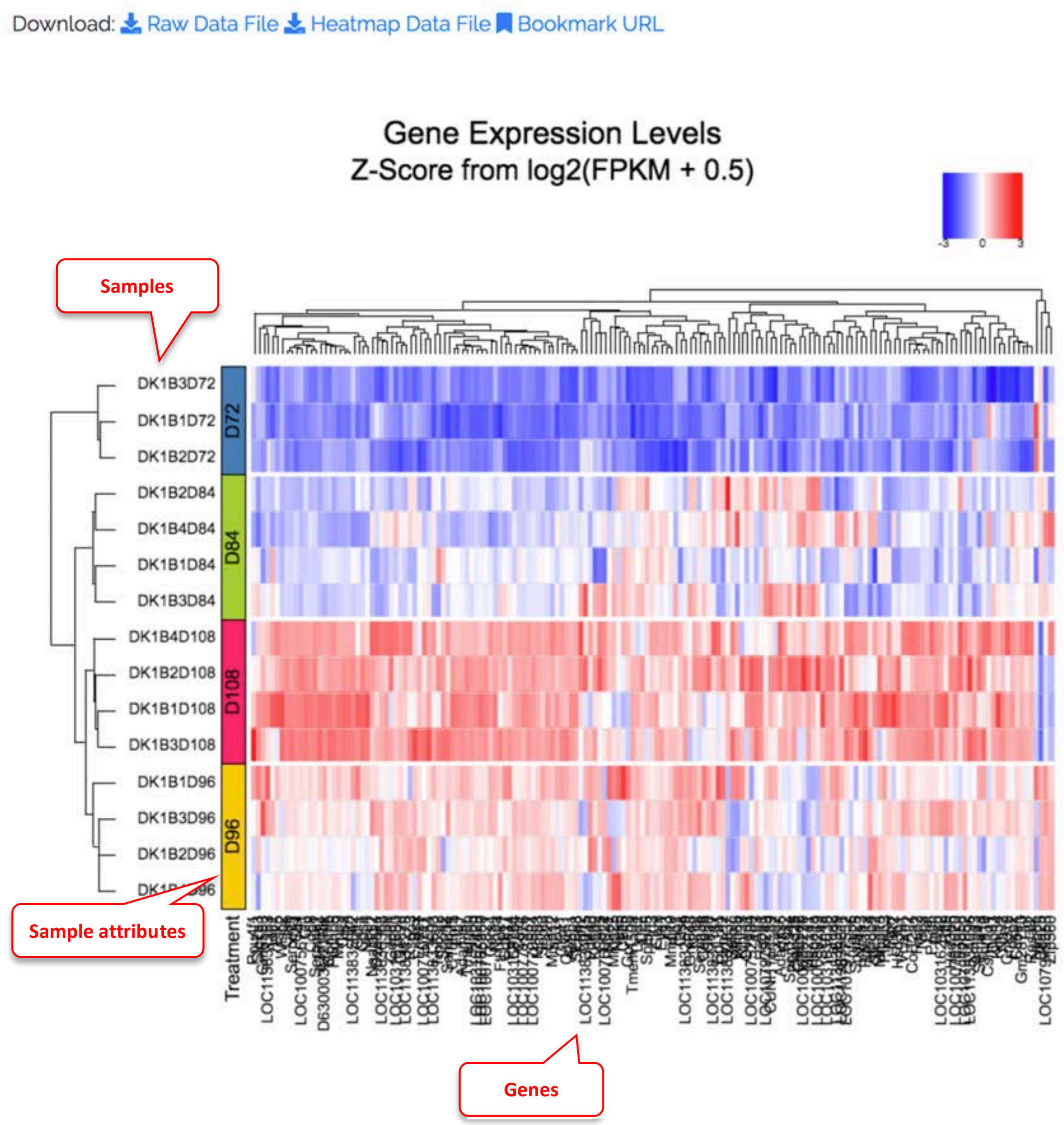
Display Options

☑ Overlay Samples

☑ Display Gene Names

☑ Display Sample IDs

The heatmap is rendered by CanvassXpress. You can change the plot size if needed.



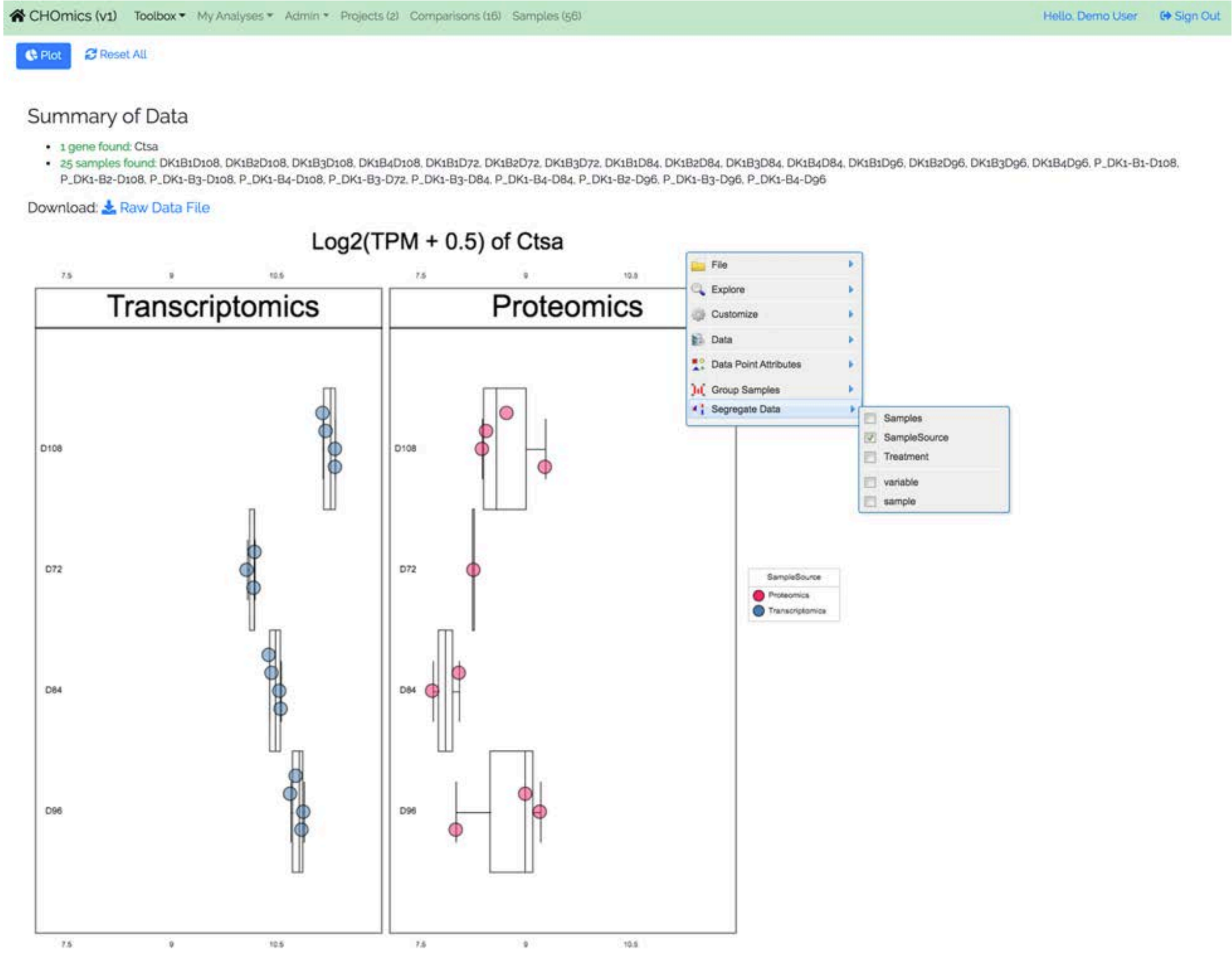
Download: Raw Data File Heatmap Data File Bookmark URL

In the example heatmap, we entered a few significantly differential expressed genes between time D72 vs time D108. From heatmap clustering, we can see that the samples are clearly clustered by time points with increase of expression on most of genes along with time.

4.1.3 Multi-omics Expression View

Besides the plotting of transcriptomics data, CHOmics also enables the visualization of other types of omics data such as proteomics, and the comparison across omics.

Here is an example of comparing gene expression (transcriptomics) and protein expression (proteomics) of gene CTSA at different time points, using the ‘Gene Expression Plot’ tool. By righ clicking the plotting area, users can group the samples by different treatment time points while segregating the data by omics type(i.e, Samplesource).



Summary of Data

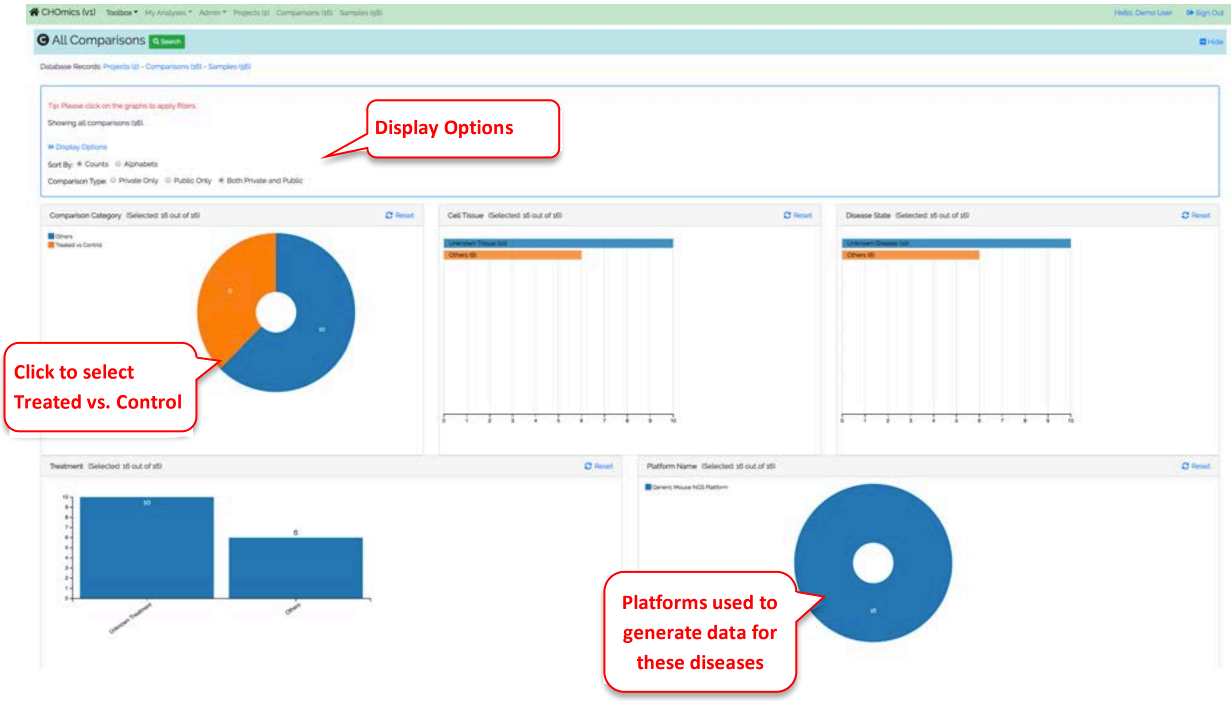
- 1 gene found: Ctsa
- 25 samples found: DK1B1D108, DK1B2D108, DK1B3D108, DK1B4D108, DK1B1D72, DK1B2D72, DK1B3D72, DK1B1D84, DK1B2D84, DK1B3D84, DK1B4D84, DK1B1D96, DK1B2D96, DK1B3D96, DK1B4D96, P_DK1-B1-D108, P_DK1-B2-D108, P_DK1-B3-D108, P_DK1-B4-D108, P_DK1-B3-D72, P_DK1-B3-D84, P_DK1-B4-D84, P_DK1-B2-D96, P_DK1-B3-D96, P_DK1-B4-D96

Download: Raw Data File

4.2 Visualize Comparison Data

4.2.1 Dashboard View of Comparison

The dashboard shows a summary of all the comparisons.



Tip: Please click on the graphs to apply filters.

Showing all comparisons (16).

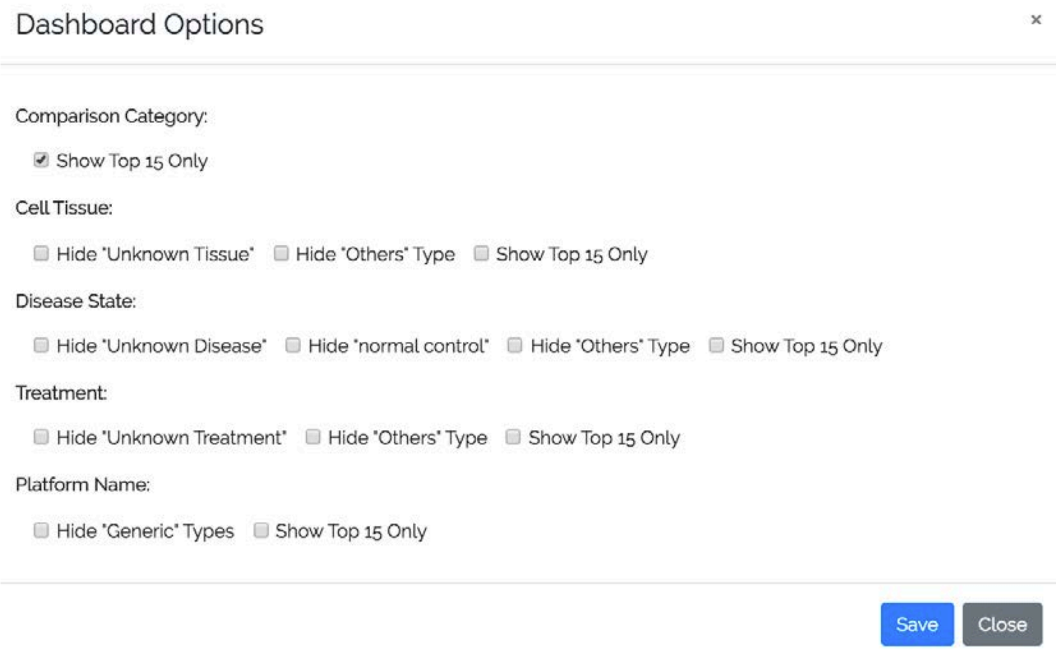
» Display Options

Sort By: ● Counts ○ Alphabets

Comparison Type: ○ Private Only ○ Public Only ● Both Private and Public

The above dashboard shows the comparisons from different Categories, Cell Type, Disease State, Treatment, Platform, etc. Below the dashboard, there is also a table listing all the comparisons.

In addition, users can set Dashboard Preference to change how the comparison summary is displayed.



Dashboard Options ×

Comparison Category:

☑ Show Top 15 Only

Cell Tissue:

☐ Hide "Unknown Tissue" ☐ Hide "Others" Type ☐ Show Top 15 Only

Disease State:

☐ Hide "Unknown Disease" ☐ Hide "normal control" ☐ Hide "Others" Type ☐ Show Top 15 Only

Treatment:

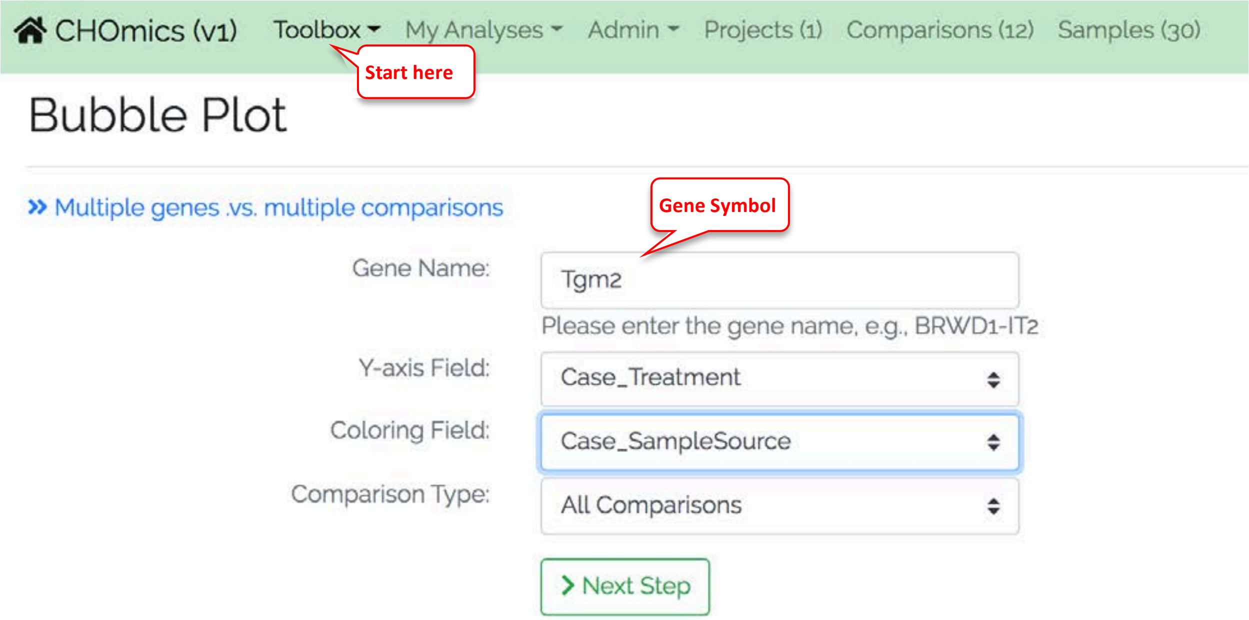
☐ Hide "Unknown Treatment" ☐ Hide "Others" Type ☐ Show Top 15 Only

Platform Name:

☐ Hide "Generic" Types ☐ Show Top 15 Only

4.2.2 Bubble Plot

Bubble plot is another useful demonstration of gene or gene set in comparisons. For each gene, you can view all the available comparisons in a bubble chart.



Start here

Bubble Plot

» Multiple genes .vs. multiple comparisons

Gene Symbol

Gene Name: Tgm2

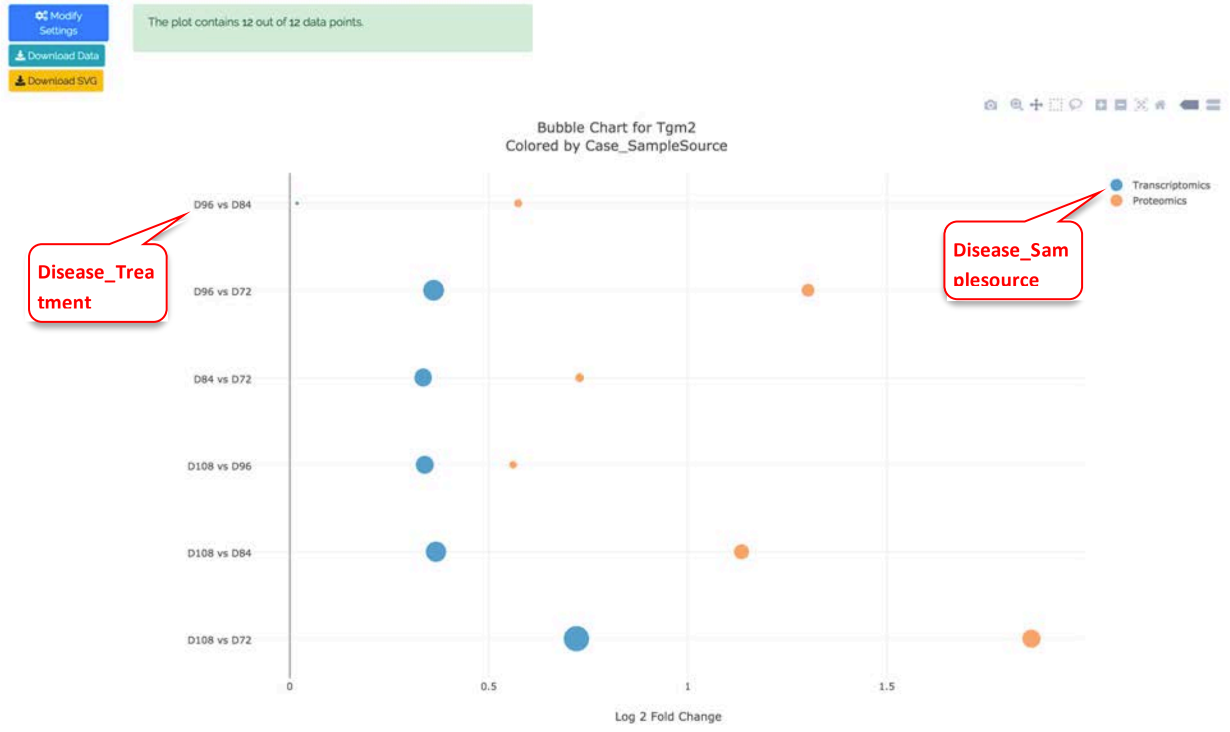
Please enter the gene name, e.g., BRWD1-IT2

Y-axis Field: Case_Treatment

Coloring Field: Case_SampleSource

Comparison Type: All Comparisons

The default settings work for most users. After clicking the Next Step button, you will see a plot like:



The plot contains 12 out of 12 data points.

Disease_Treatment

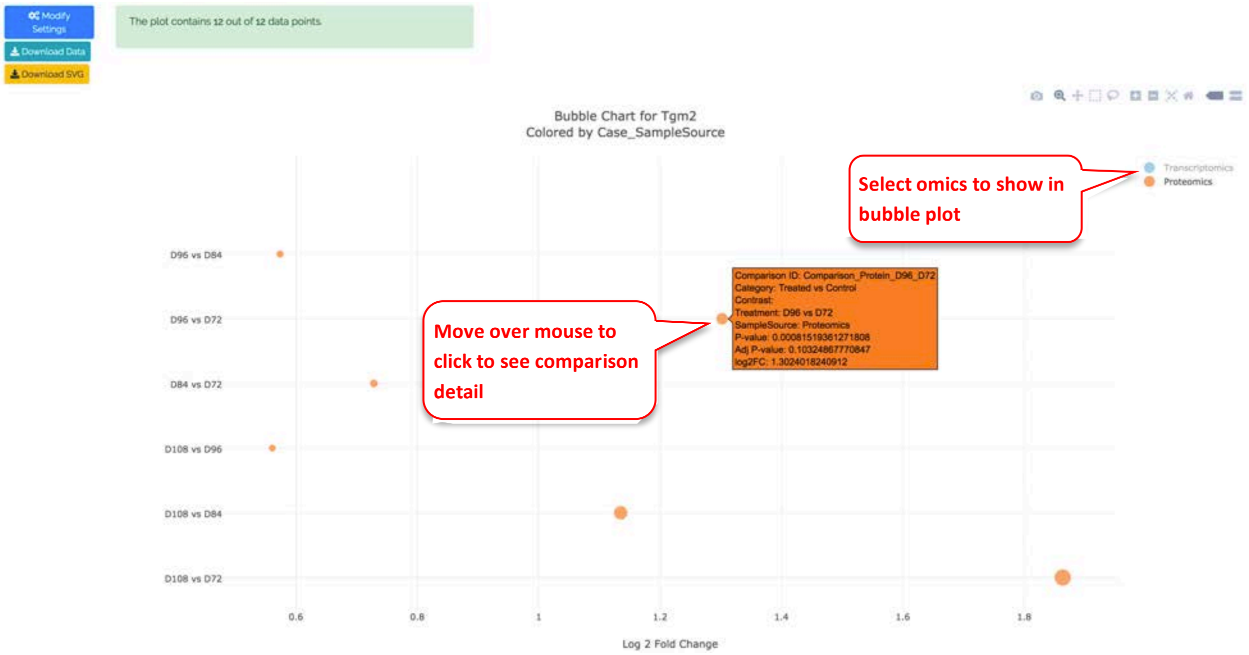
Disease_Samplesource

In the bubble plot, the X-axis shows log2 Fold Change of the comparison, the Y-axis shows 'Case_treatment'. Each dot represents the comparison result of this gene from one comparison. The color of the dot represent 'Case_Samplesource' (i.e, here we set as omics type), and the size of the dot represent significance (-log10(FDR), larger is more significant).

The user can click and unclick the color legend at right to select or deselect omics types. When mouse over a dot, more details are shown. And the user can also click the dot to link to other graphs.

The tool bars at top right corner allows the user to zoom and pan the graph.

The screenshot below shows the same bubble chart after selecting one omics type (i.e,transcriptomics), and zoom into a portion of the chart.



The plot contains 12 out of 12 data points.

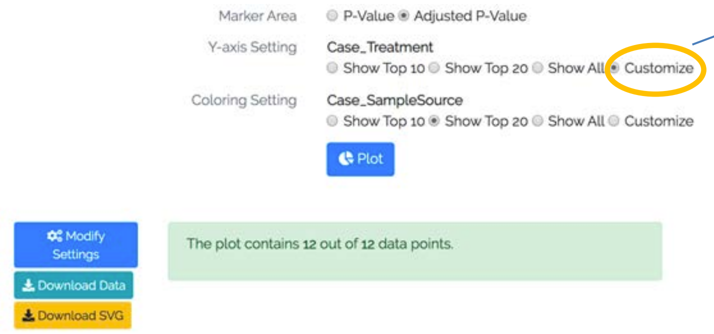
Comparison ID: Comparison_Protein_D96_D72
Category: Treated vs Control
Contrast:
Treatment: D96 vs D72
SampleSource: Proteomics
P-value: 0.00081519361271808
Adj P-value: 0.10324867770847
log2FC: 1.3024018240912

Select omics to show in bubble plot

Move over mouse to click to see comparison detail

Data Filter and Advanced Settings in Bubble Plot

In addition, advanced users can change settings by click "Modify Settings Button". For example, the user may want to show a selected list of diseases. After clicking Customize in Case_Treatment, user can select which treatments to display in the pop-up window.

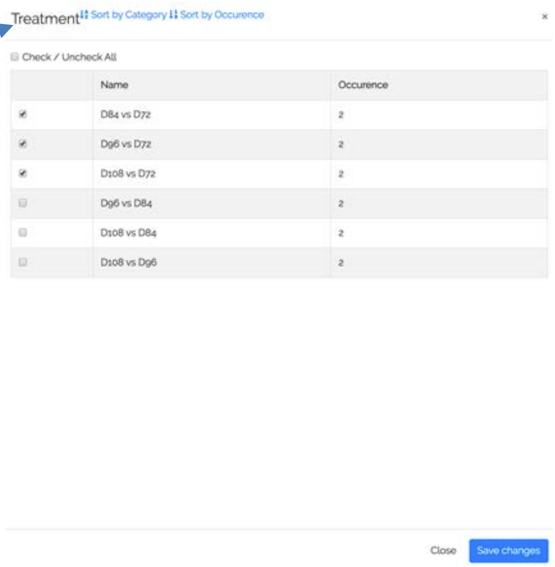


Marker Area ○ P-Value ● Adjusted P-Value

Y-axis Setting Case_Treatment
○ Show Top 10 ○ Show Top 20 ○ Show All ● Customize

Coloring Setting Case_SampleSource
○ Show Top 10 ● Show Top 20 ○ Show All ○ Customize

The plot contains 12 out of 12 data points.

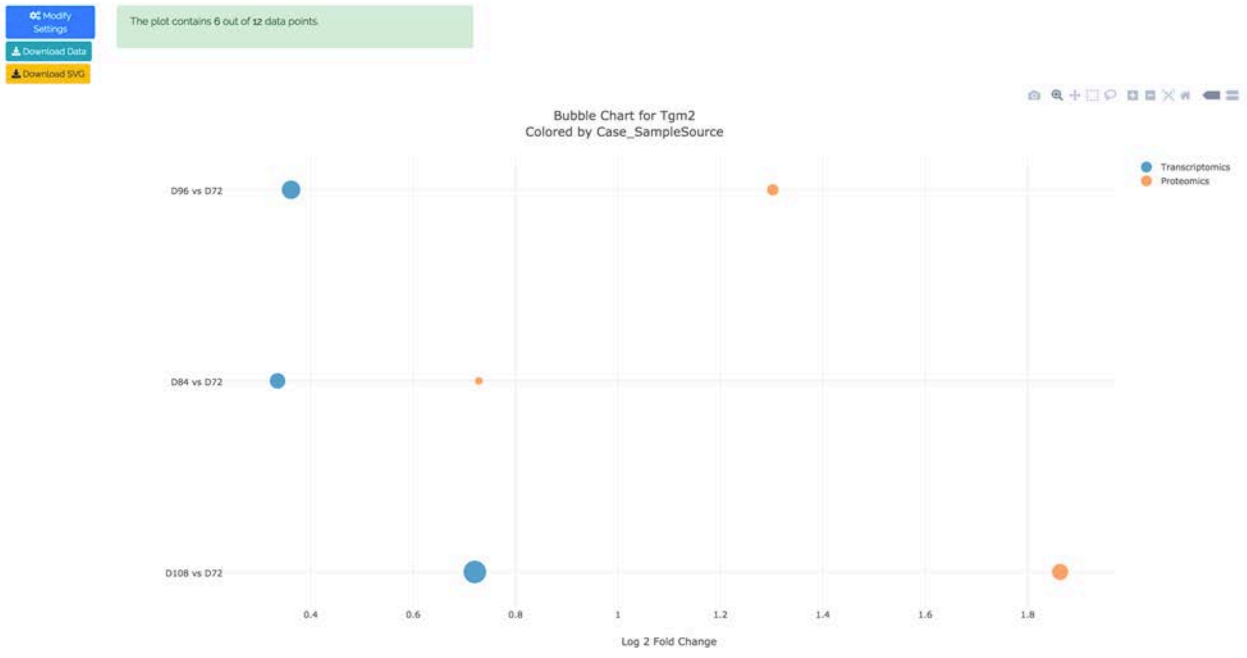


Treatment Sort by Category Sort by Occurence

□ Check / Uncheck All

	Name	Occurence
☑	D84 vs D72	2
☑	D96 vs D72	2
☑	D108 vs D72	2
□	D96 vs D84	2
□	D108 vs D84	2
□	D108 vs D96	2

After modifying the setting, the user can click plot button to view the new chart. The system will display how many data points are chosen based on the filter.

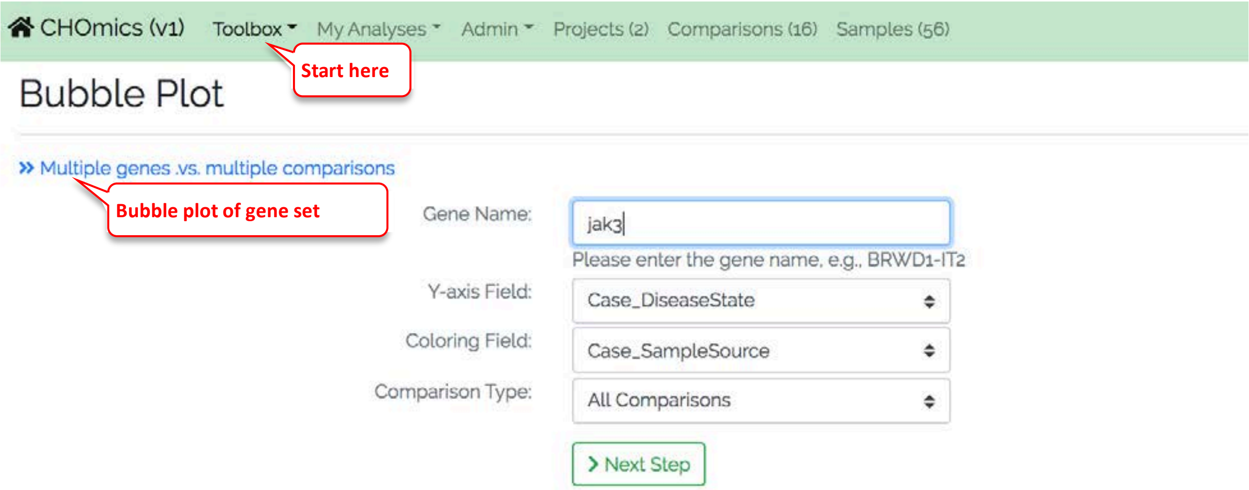


The plot contains 6 out of 12 data points.

Bubble Plot of Multiple Genes and Multiple Comparisons

It can be useful to look at a set of genes (e.g. all differentially expressed genes, or genes from a certain pathways) in a set of related comparisons (e.g. all from the same disease).

To view this type of bubble plot, select the link for Multiple Genes vs. multiple comparisons.



Start here

Bubble plot of gene set

Gene Name: jak3

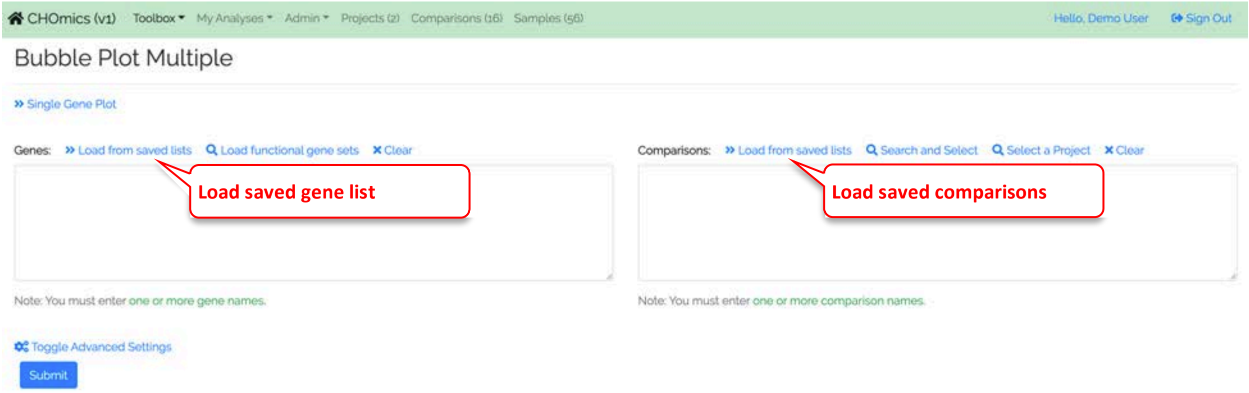
Please enter the gene name, e.g., BRWD1-IT2

Y-axis Field: Case_DiseaseState

Coloring Field: Case_SampleSource

Comparison Type: All Comparisons

In the Genes and Comparisons Bubble plot window, you can now enter the symbols of the genes, and the comparison names. However, it is much easier to use the saved genes and saved comparisons features, or other tools from the system to quickly get a get set. Please see below for details.



Load saved gene list

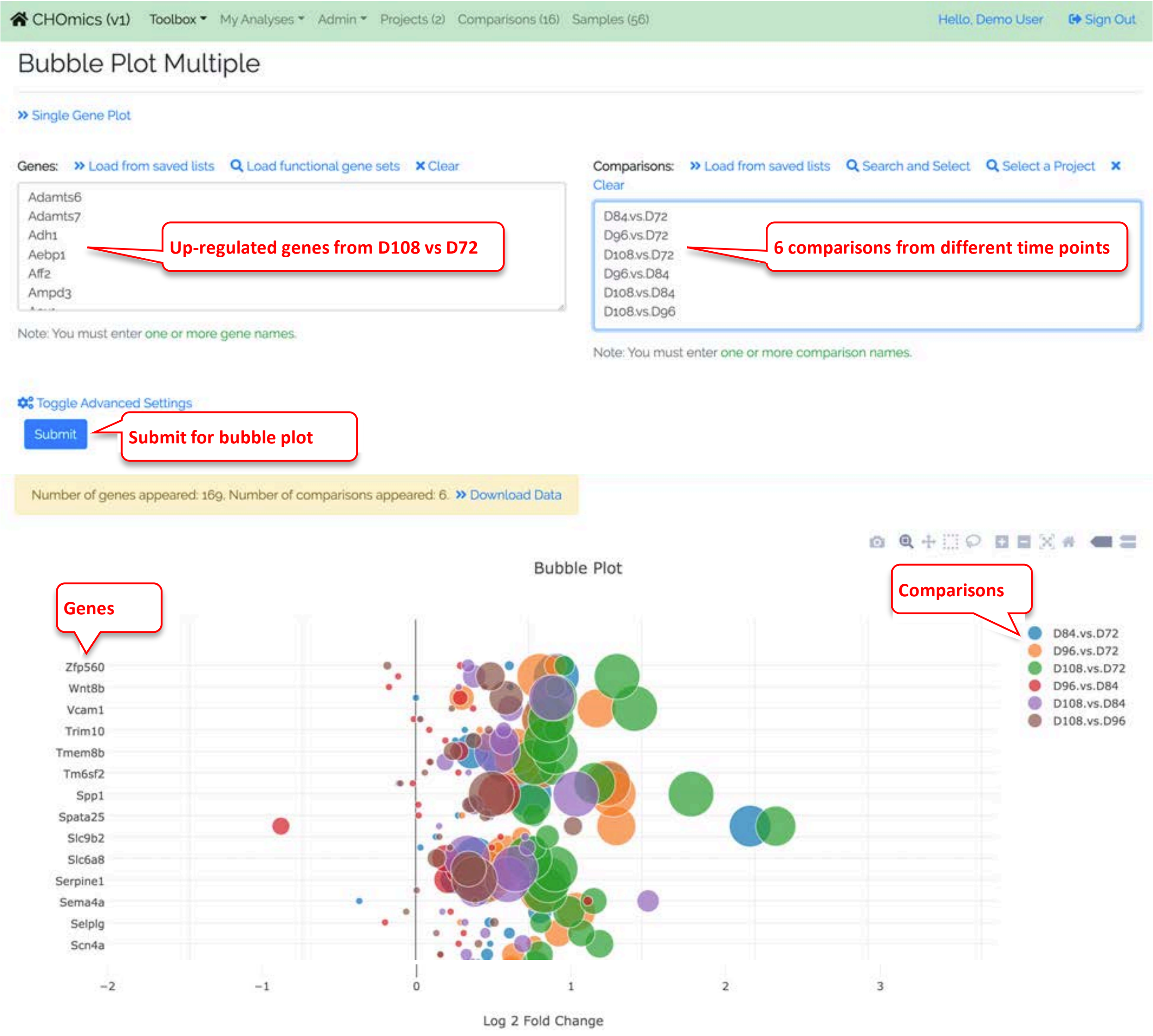
Load saved comparisons

Note: You must enter one or more gene names.

Note: You must enter one or more comparison names.

In the example below, we use dashboard to select 6 comparisons that are for different time points in CHO cell lines. We save the comparisons and load in the bubble plot tool. For gene list, we get the up-regulated genes from comparison D72 vs D108, and paste into the gene names fields.

In the bubble plot, the gene symbols are listed in Y-axis. The X-axis represents logFC, and color of the bubble represents comparison; the size of the bubble represents the significance.



CHOmics (v1) Toolbox ▾ My Analyses ▾ Admin ▾ Projects (2) Comparisons (16) Samples (56) Hello, Demo User Sign Out

Bubble Plot Multiple

» Single Gene Plot

Genes: » Load from saved lists Load functional gene sets ✕ Clear

Adamts6
Adamts7
Adh1
Aebp1
Aff2
Ampd3

Note: You must enter one or more gene names.

Up-regulated genes from D108 vs D72

Comparisons: » Load from saved lists Search and Select Select a Project ✕ Clear

D84.vs.D72
D96.vs.D72
D108.vs.D72
D96.vs.D84
D108.vs.D84
D108.vs.D96

Note: You must enter one or more comparison names.

6 comparisons from different time points

Toggle Advanced Settings

Submit

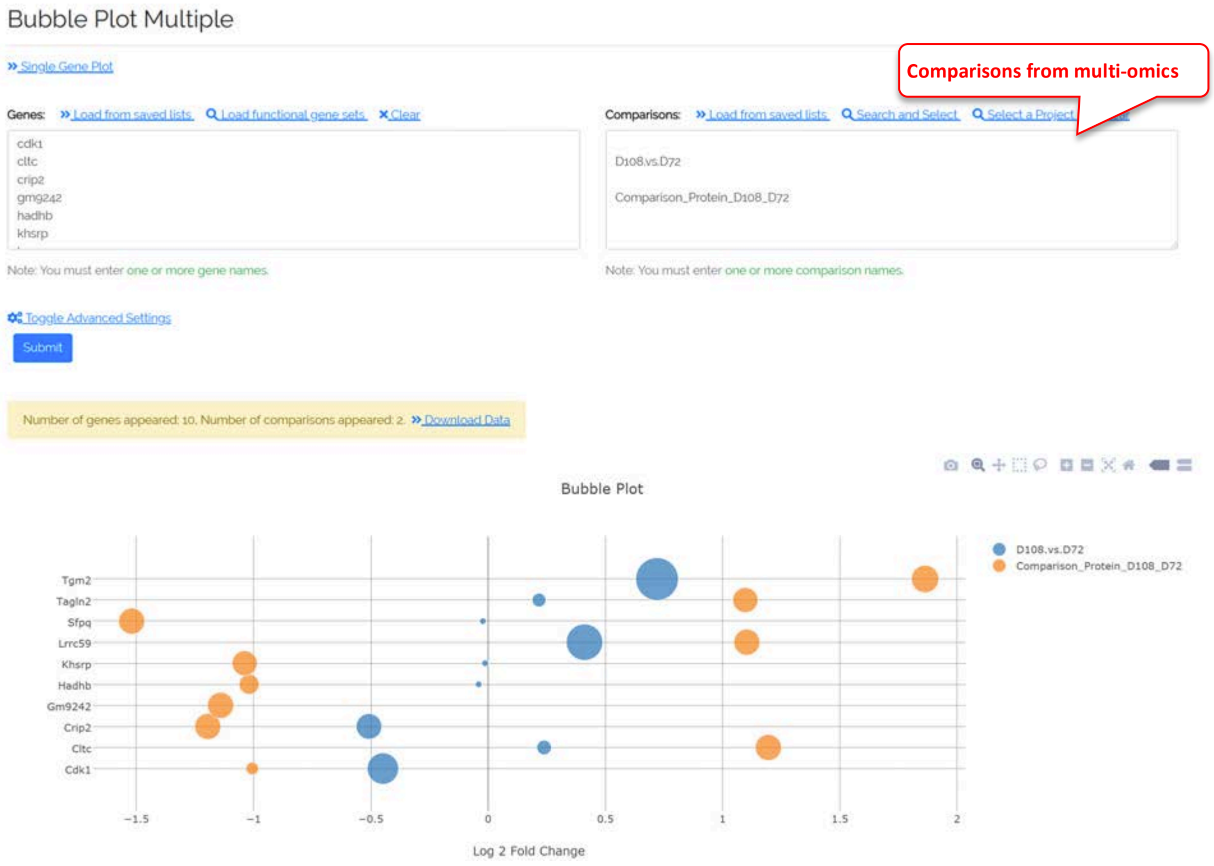
Submit for bubble plot

Number of genes appeared: 169, Number of comparisons appeared: 6. » Download Data

In the legend, the color keys for comparisons are shown. You can click the color key in the legend to hide/show comparisons. The size of the color dot in the legend correlates to the largest bubble for that comparison, which is the most significant gene with the smallest FDR.

Bubble Plot of Multiple Omics data

Similar to the bubble plot of multiple genes across multiple comparisons, users can further compare the genes on the comparisons from different omics data.



Bubble Plot Multiple

» Single Gene Plot

Comparisons from multi-omics

Genes: » Load from saved lists Load functional gene sets ✕ Clear

cdk1
cltc
crip2
gm9242
hadhb
khsrp

Note: You must enter one or more gene names.

Comparisons: » Load from saved lists Search and Select Select a Project

D108.vs.D72

Comparison_Protein_D108_D72

Note: You must enter one or more comparison names.

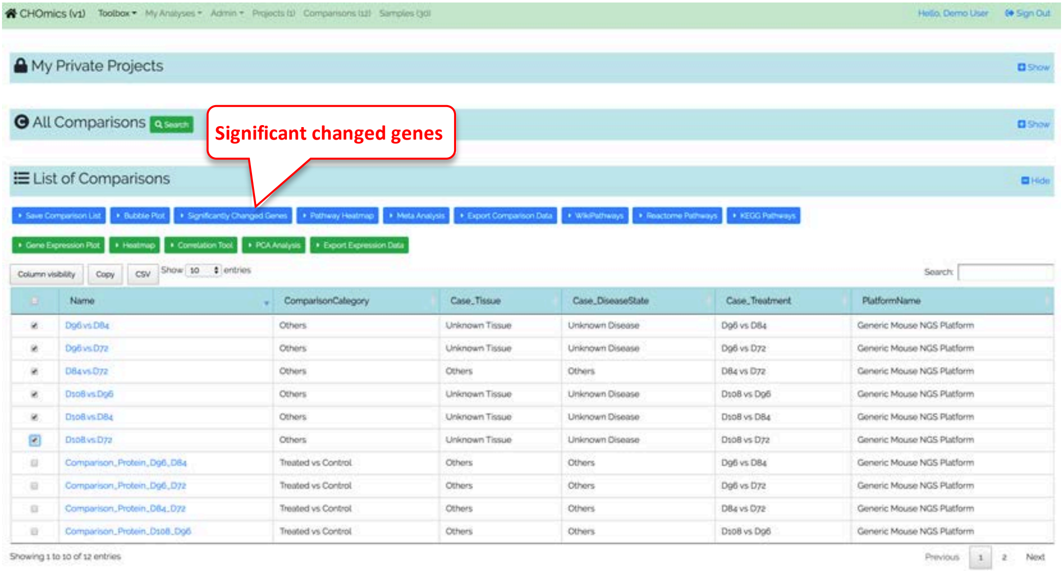
Toggle Advanced Settings

Submit

Number of genes appeared: 10. Number of comparisons appeared: 2. » Download Data

4.2.3 Get significant genes from comparisons

Another way to get a gene set to visualize in the genes/comparisons bubble plot is to filter for significantly changed genes. To do this, first select a few comparisons from the dash board, and click the "View Significantly Changed Genes" button.



My Private Projects Show

All Comparisons Search Show

Significant changed genes

List of Comparisons Hide

Show 10 entries Search:

□	Name	ComparisonCategory	Case_Tissue	Case_DiseaseState	Case_Treatment	PlatformName
☑	D96.vs.D84	Others	Unknown Tissue	Unknown Disease	D96 vs D84	Generic Mouse NGS Platform
☑	D96.vs.D72	Others	Unknown Tissue	Unknown Disease	D96 vs D72	Generic Mouse NGS Platform
☑	D84.vs.D72	Others	Others	Others	D84 vs D72	Generic Mouse NGS Platform
☑	D108.vs.D96	Others	Unknown Tissue	Unknown Disease	D108 vs D96	Generic Mouse NGS Platform
☑	D108.vs.D84	Others	Unknown Tissue	Unknown Disease	D108 vs D84	Generic Mouse NGS Platform
☑	D108.vs.D72	Others	Unknown Tissue	Unknown Disease	D108 vs D72	Generic Mouse NGS Platform
□	Comparison_Protein_D96_D84	Treated vs Control	Others	Others	D96 vs D84	Generic Mouse NGS Platform
□	Comparison_Protein_D96_D72	Treated vs Control	Others	Others	D96 vs D72	Generic Mouse NGS Platform
□	Comparison_Protein_D84_D72	Treated vs Control	Others	Others	D84 vs D72	Generic Mouse NGS Platform
□	Comparison_Protein_D108_D96	Treated vs Control	Others	Others	D108 vs D96	Generic Mouse NGS Platform

Showing 1 to 10 of 12 entries Previous 1 2 Next

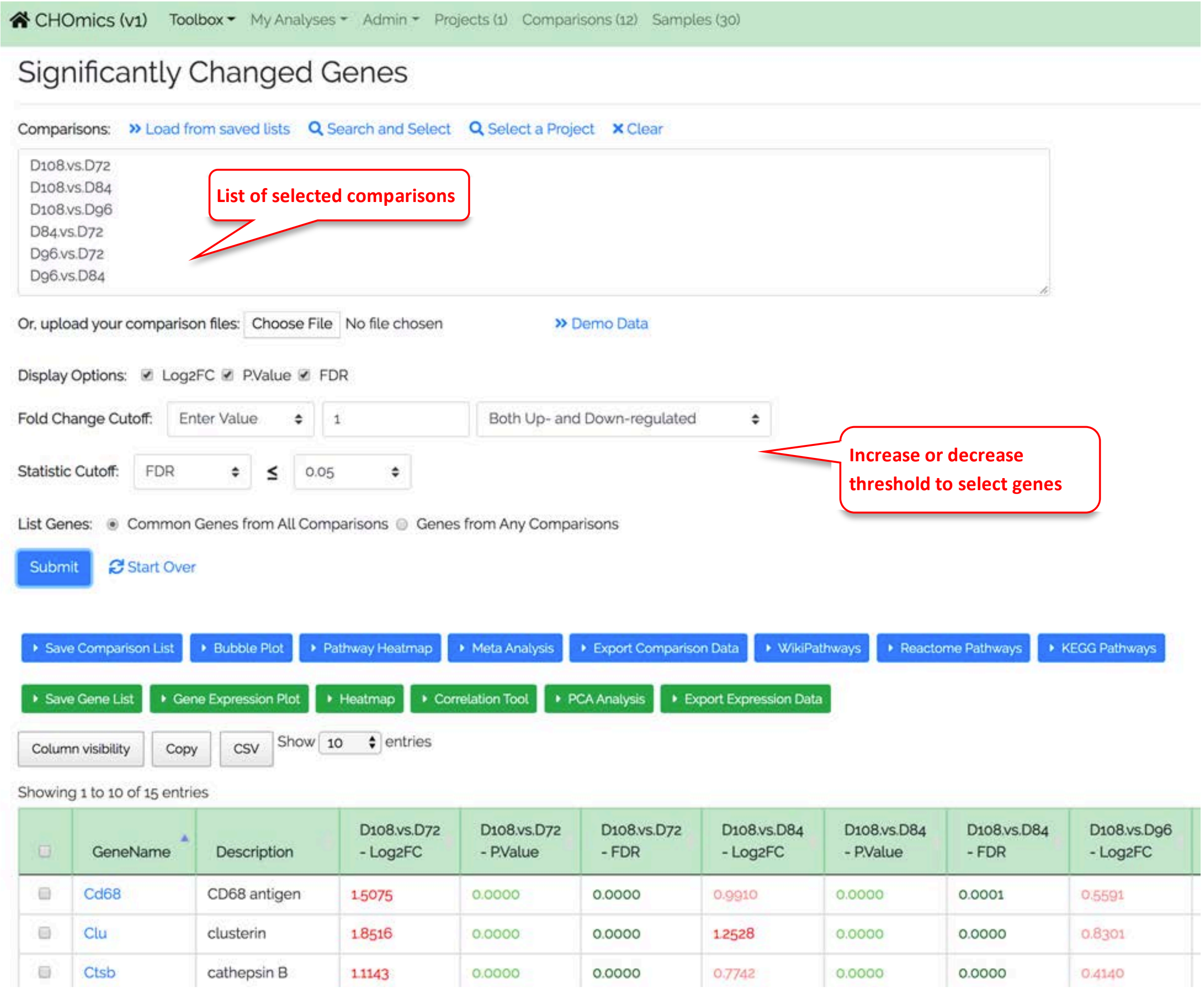
Dashboard filter.

In table, select comparisons and view significantly changed genes.

In the significantly Changed Genes window, the comparisons from the previous page are already loaded. You can add or remove comparisons if needed.

Now select direction (up-, down-, or both), and use the logFC cutoff and FDR value to get a list of genes. Depending on the comparisons, sometimes you may need to adjust the logFC and FDR values to get a good list of genes. In general, for bubble plot, using <100 genes will make the graph easier to read.

Once you are happy with the gene list, you can save it. You can also export the list for later use.



Significantly Changed Genes

Comparisons:

D108.vs.D72
D108.vs.D84
D108.vs.D96
D84.vs.D72
D96.vs.D72
D96.vs.D84

List of selected comparisons

Or, upload your comparison files: Choose File No file chosen

Display Options: ☑ Log2FC ☑ P.Value ☑ FDR

Fold Change Cutoff: Enter Value 1 Both Up- and Down-regulated

Statistic Cutoff: FDR ≤ 0.05

Increase or decrease threshold to select genes

List Genes: ● Common Genes from All Comparisons ○ Genes from Any Comparisons

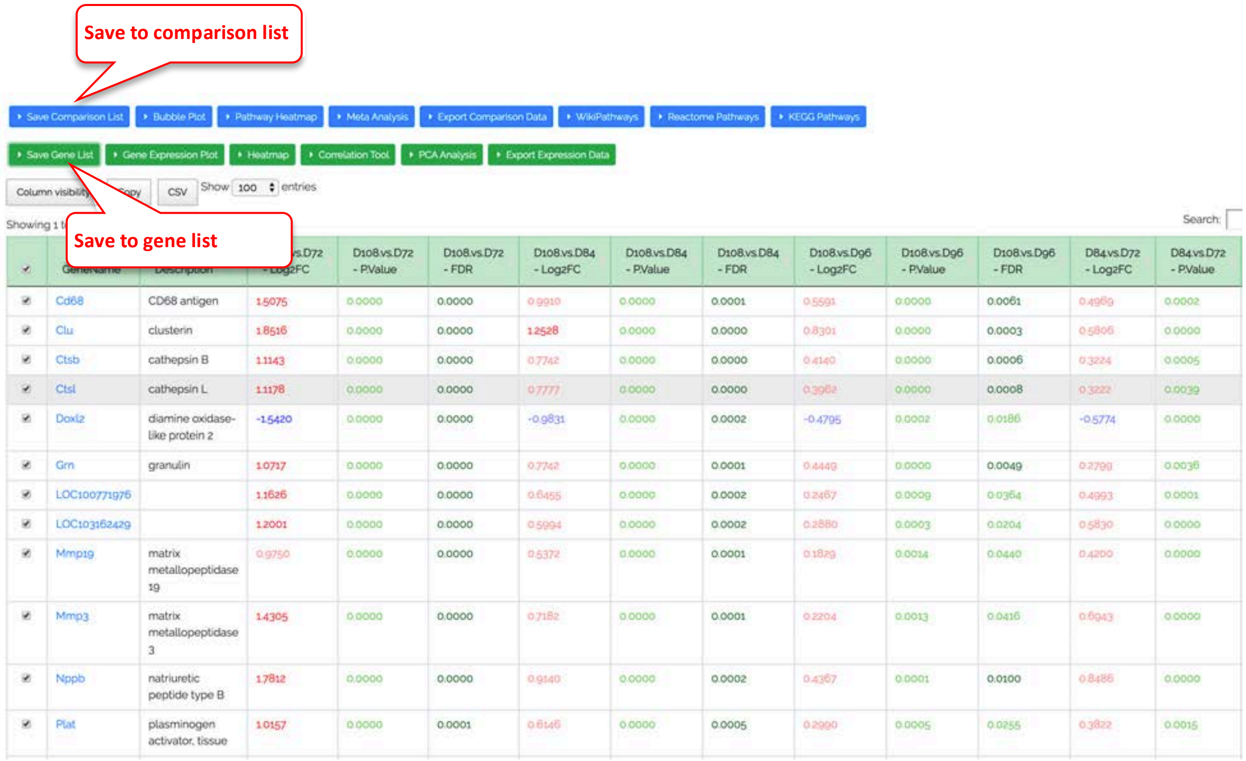
Show 10 entries

Showing 1 to 10 of 15 entries

□	GeneName	Description	D108.vs.D72 - Log2FC	D108.vs.D72 - P.Value	D108.vs.D72 - FDR	D108.vs.D84 - Log2FC	D108.vs.D84 - P.Value	D108.vs.D84 - FDR	D108.vs.D96 - Log2FC
□	Cd68	CD68 antigen	1.5075	0.0000	0.0000	0.9910	0.0000	0.0001	0.5591
□	Clu	clusterin	1.8516	0.0000	0.0000	1.2528	0.0000	0.0000	0.8301
□	Ctsb	cathepsin B	1.1143	0.0000	0.0000	0.7742	0.0000	0.0000	0.4140

View Significantly Changed Genes in Bubble Plot

Back to the bubble plot, you can load the saved comparisons and saved genes and view the plot.

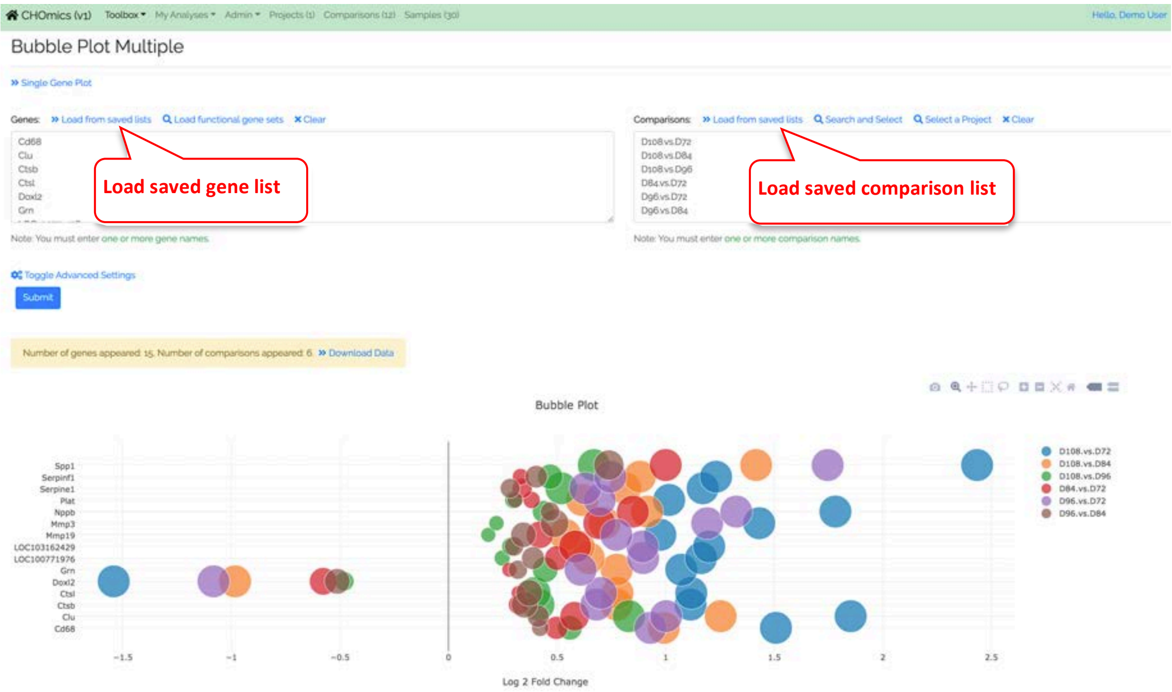


Show 100 entries

Showing 1 t

☑	GeneName	Description	D108.vs.D72 - Log2FC	D108.vs.D72 - P.Value	D108.vs.D72 - FDR	D108.vs.D84 - Log2FC	D108.vs.D84 - P.Value	D108.vs.D84 - FDR	D108.vs.D96 - Log2FC	D108.vs.D96 - P.Value	D108.vs.D96 - FDR	D84.vs.D72 - Log2FC	D84.vs.D72 - P.Value
☑	Cd68	CD68 antigen	1.5075	0.0000	0.0000	0.9910	0.0000	0.0001	0.5591	0.0000	0.0061	0.4969	0.0002
☑	Clu	clusterin	1.8516	0.0000	0.0000	1.2528	0.0000	0.0000	0.8301	0.0000	0.0003	0.5806	0.0000
☑	Ctsb	cathepsin B	1.1143	0.0000	0.0000	0.7742	0.0000	0.0000	0.4140	0.0000	0.0006	0.3224	0.0005
☑	Ctsl	cathepsin L	1.1178	0.0000	0.0000	0.7777	0.0000	0.0000	0.3962	0.0000	0.0008	0.3222	0.0039
☑	Doxl2	diamine oxidase-like protein 2	-1.5420	0.0000	0.0000	-0.9831	0.0000	0.0002	-0.4795	0.0002	0.0186	-0.5774	0.0000
☑	Grn	granulin	1.0717	0.0000	0.0000	0.7742	0.0000	0.0001	0.4449	0.0000	0.0049	0.2799	0.0036
☑	LOC100771976		1.1626	0.0000	0.0000	0.6455	0.0000	0.0002	0.2467	0.0009	0.0364	0.4993	0.0001
☑	LOC103162429		1.2001	0.0000	0.0000	0.5994	0.0000	0.0002	0.2880	0.0003	0.0204	0.5830	0.0000
☑	Mmp19	matrix metallopeptidase 19	0.9750	0.0000	0.0000	0.5372	0.0000	0.0001	0.1829	0.0014	0.0440	0.4200	0.0000
☑	Mmp3	matrix metallopeptidase 3	1.4305	0.0000	0.0000	0.7182	0.0000	0.0001	0.2204	0.0013	0.0416	0.6943	0.0000
☑	Nppb	natriuretic peptide type B	1.7812	0.0000	0.0000	0.9140	0.0000	0.0002	0.4367	0.0001	0.0100	0.8486	0.0000
☑	Plat	plasminogen activator, tissue	1.0157	0.0000	0.0001	0.6146	0.0000	0.0005	0.2990	0.0005	0.0255	0.3822	0.0015

In the example below, it can be seen that most significant genes come from down-regulated direction from the first four comparisons.



Bubble Plot Multiple

Genes:

```
Cd68
Clu
Ctsb
Ctsl
Doxl2
Grn
```

Note: You must enter one or more gene names.

Comparisons:

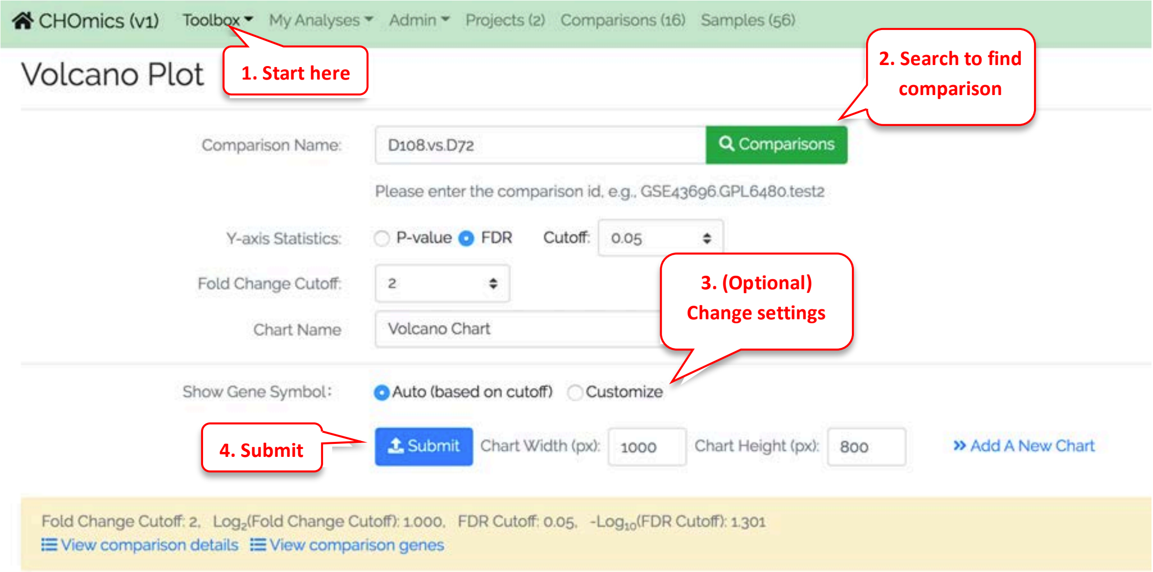
```
D108.vs.D72
D108.vs.D84
D108.vs.D96
D84.vs.D72
D96.vs.D72
D96.vs.D84
```

Note: You must enter one or more comparison names.

Number of genes appeared: 15. Number of comparisons appeared: 6.

4.2.4 Volcano Plot

Volcano plot is useful to view a top level summary of how many genes are significantly up- or down-regulated in a comparison.



Volcano Plot

1. Start here

2. Search to find comparison

Comparison Name: D108.vs.D72 [Comparisons]

Please enter the comparison id, e.g., GSE43696.GPL6480.test2

Y-axis Statistics: ○ P-value ● FDR Cutoff: 0.05

Fold Change Cutoff: 2

Chart Name: Volcano Chart

3. (Optional) Change settings

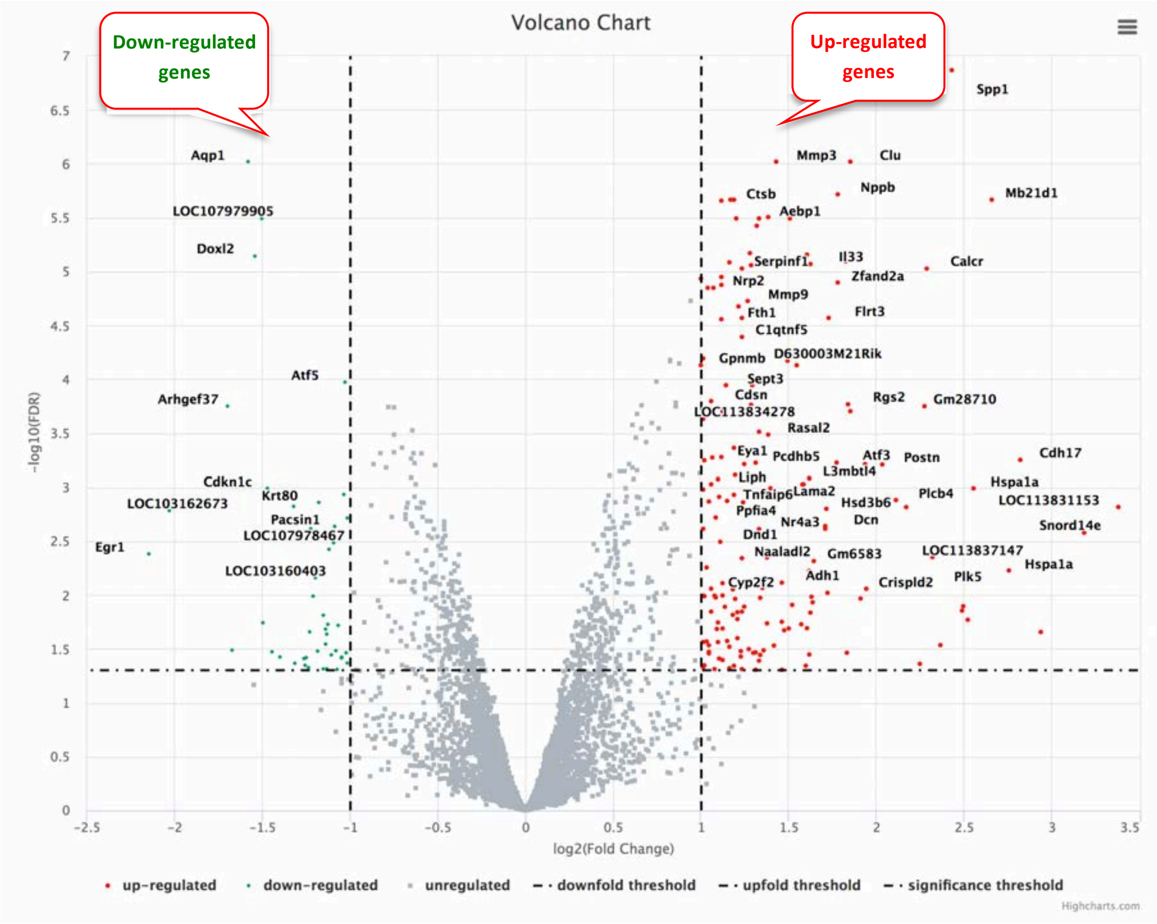
Show Gene Symbol: ● Auto (based on cutoff) ○ Customize

4. Submit

[Submit] Chart Width (px): 1000 Chart Height (px): 800 » Add A New Chart

Fold Change Cutoff: 2, Log_2(Fold Change Cutoff): 1.000, FDR Cutoff: 0.05, $-\text{Log}_{10}$(FDR Cutoff): 1.301

View comparison details View comparison genes

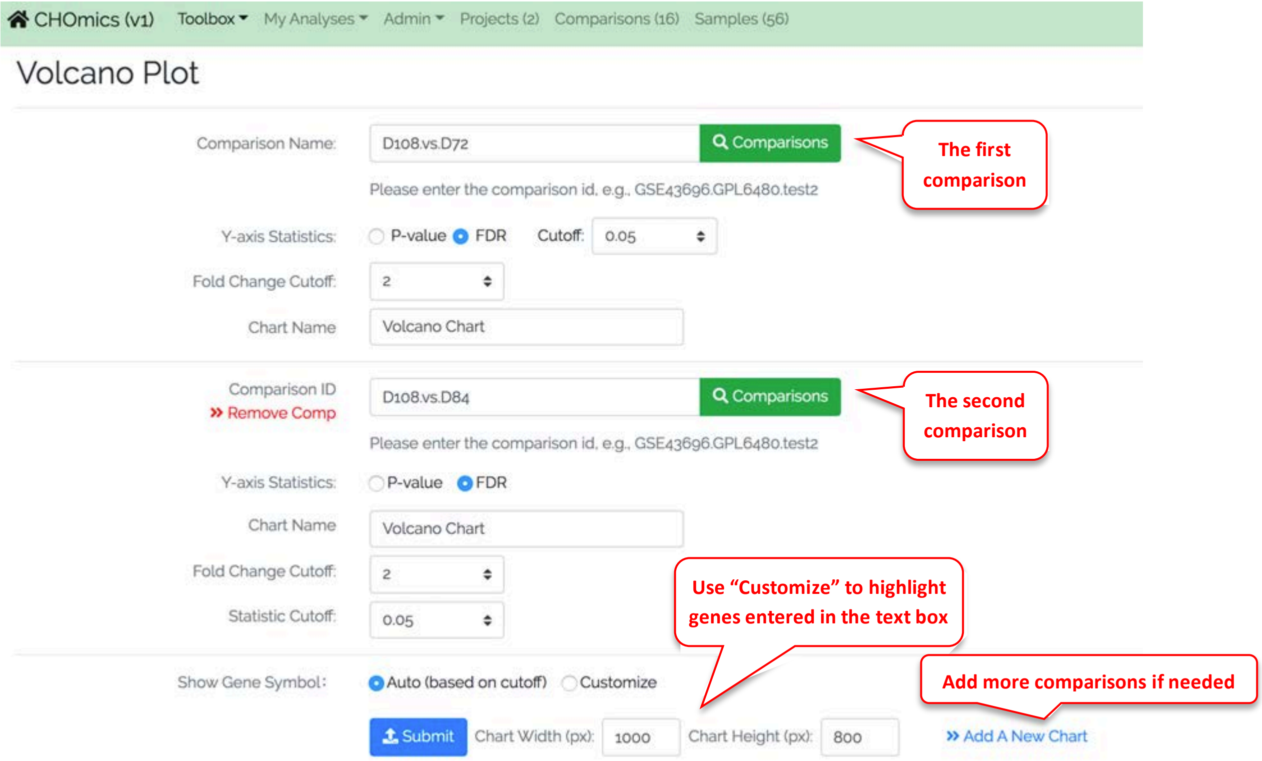


You can use mouse to drag over an area to zoom in.

Mouse over a point will show the gene details. Click the data point will show you links to other graphs.

View Multiple Volcano Plots Together

Users can also show multiple comparisons side-by-side. If needed, the user can also highlight the same group of genes across the volcano plots.



Volcano Plot

Comparison Name: D108.vs.D72 Comparisons

Please enter the comparison id, e.g., GSE43696.GPL6480.test2

Y-axis Statistics: ○ P-value ● FDR Cutoff: 0.05

Fold Change Cutoff: 2

Chart Name: Volcano Chart

Comparison ID » Remove Comp: D108.vs.D84 Comparisons

Please enter the comparison id, e.g., GSE43696.GPL6480.test2

Y-axis Statistics: ○ P-value ● FDR

Chart Name: Volcano Chart

Fold Change Cutoff: 2

Statistic Cutoff: 0.05

Show Gene Symbol: ● Auto (based on cutoff) ○ Customize

Submit Chart Width (px): 1000 Chart Height (px): 800 » Add A New Chart

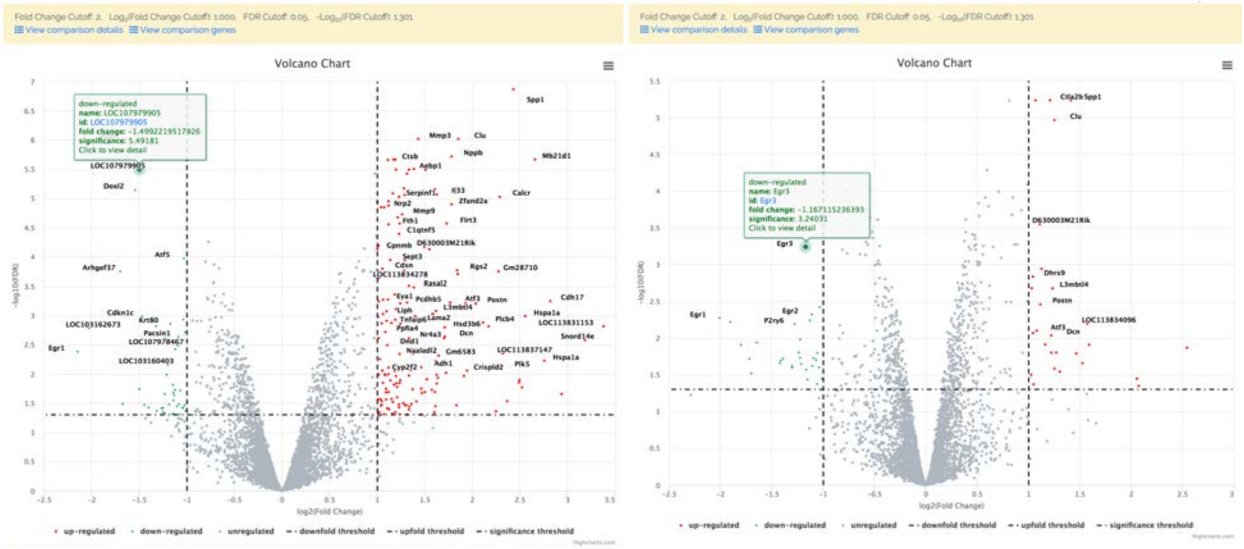
The first comparison

The second comparison

Use “Customize” to highlight genes entered in the text box

Add more comparisons if needed

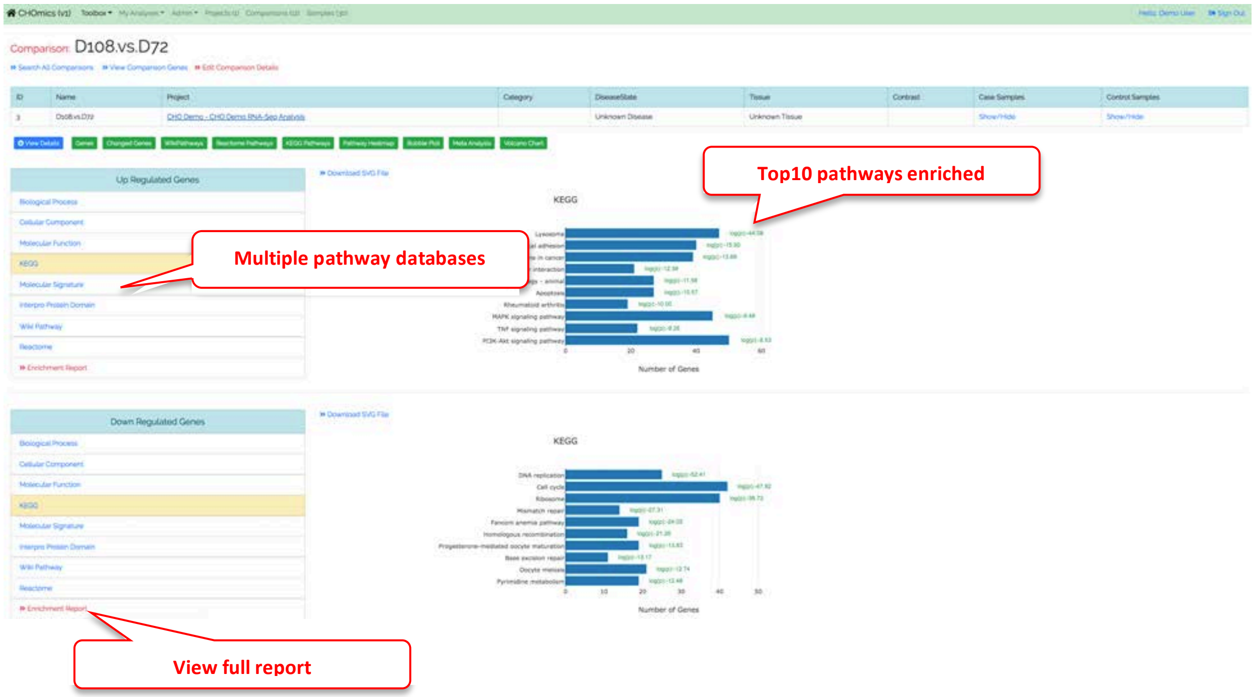
The resulting volcano plots are shown as below. Selected genes are shown as orange dots.



4.3 Visualize functional pathway

4.3.1 Enrichment from Up and Down Regulated Genes

When you view details of a comparison, the functional enrichment results are shown. Briefly, for each comparison, we generated the up- and down- regulate gene lists, and use these lists to compare with all genes in the genome to identify functions that are significantly enriched.



Comparison: D108.vs.D72

ID	Name	Project	Category	DiseaseState	Tissue	Contrast	Case Samples	Control Samples
3	D108.vs.D72	CHO Demo - CHO Demo RNA-Seq Analysis		Unknown Disease	Unknown Tissue		Show/Hide	Show/Hide

Top10 pathways enriched

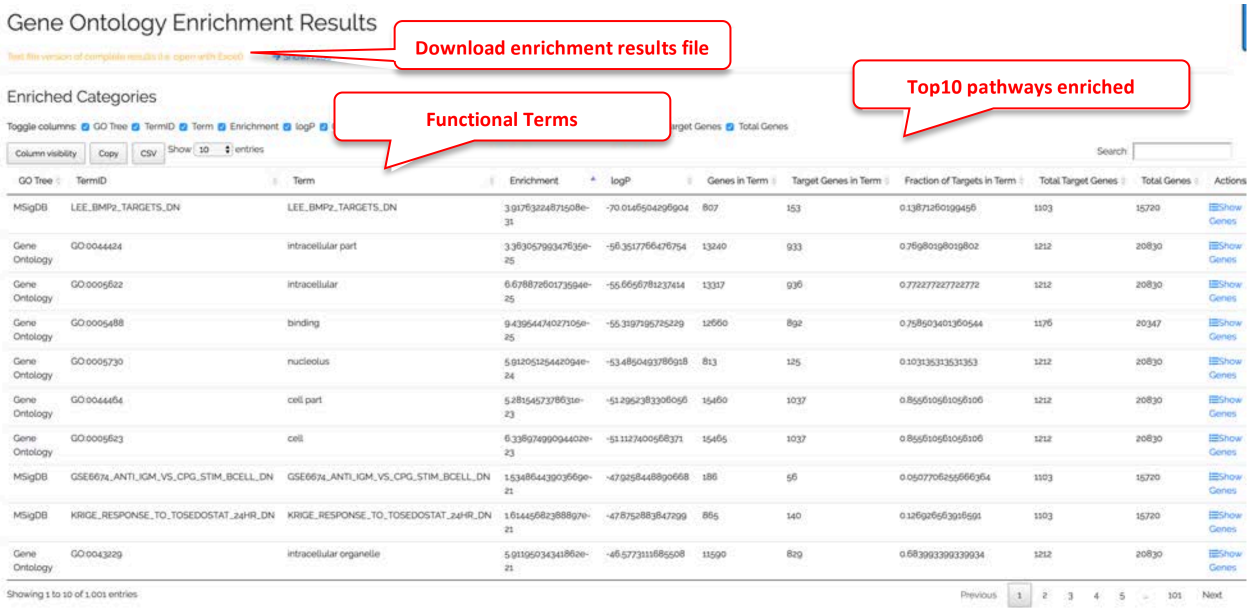
Multiple pathway databases

View full report

In the example above, this comparison is between D108 vs D72, and the top up-regulated biological processed are response to virus, immune effector process.

Click the left menu will switch the bar charts for different categories (Gene Ontology, KEGG, Molecular signature, Protein domain etc).

The bar charts here show the top 10 categories. To view complete results, click the Enrichment Report.



Gene Ontology Enrichment Results

Text file version of complete results (i.e. open with Excel)

Download enrichment results file

Enriched Categories

Functional Terms

Top10 pathways enriched

Toggle columns: ☑ GO Tree ☑ TermID ☑ Term ☑ Enrichment ☑ logP ☑ … Target Genes ☑ Total Genes

Column visibility | Copy | CSV | Show 10 entries | Search:

GO Tree	TermID	Term	Enrichment	logP	Genes in Term	Target Genes in Term	Fraction of Targets in Term	Total Target Genes	Total Genes	Actions
MSigDB	LEE_BMP2_TARGETS_DN	LEE_BMP2_TARGETS_DN	3.91763224871508e-31	-70.0146504296904	807	153	0.13871260199456	1103	15720	☰Show Genes
Gene Ontology	GO:0044424	intracellular part	3.36305799347635e-25	-56.3517766476754	13240	933	0.769801980198O2	1212	20830	☰Show Genes
Gene Ontology	GO:0005622	intracellular	6.67887260173594e-25	-55.6656781237414	13317	936	0.772277227722772	1212	20830	☰Show Genes
Gene Ontology	GO:0005488	binding	9.43954474027105e-25	-55.3197195725229	12660	892	0.758503401360544	1176	20347	☰Show Genes
Gene Ontology	GO:0005730	nucleolus	5.91205125442094e-24	-53.4850493786918	813	125	0.103135313531353	1212	20830	☰Show Genes
Gene Ontology	GO:0044464	cell part	5.2815457378631e-23	-51.2952383306056	15460	1037	0.855610561056106	1212	20830	☰Show Genes
Gene Ontology	GO:0005623	cell	6.33897499094402e-23	-51.1127400568371	15465	1037	0.855610561056106	1212	20830	☰Show Genes
MSigDB	GSE6674_ANTI_IGM_VS_CPG_STIM_BCELL_DN	GSE6674_ANTI_IGM_VS_CPG_STIM_BCELL_DN	1.53486443903669e-21	-47.9258448890668	186	56	0.0507706255666364	1103	15720	☰Show Genes
MSigDB	KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_DN	KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_DN	1.61445682388897e-21	-47.8752883847299	865	140	0.126926563916591	1103	15720	☰Show Genes
Gene Ontology	GO:0043229	intracellular organelle	5.91195034341862e-21	-46.5773111685508	11590	829	0.683993399339934	1212	20830	☰Show Genes

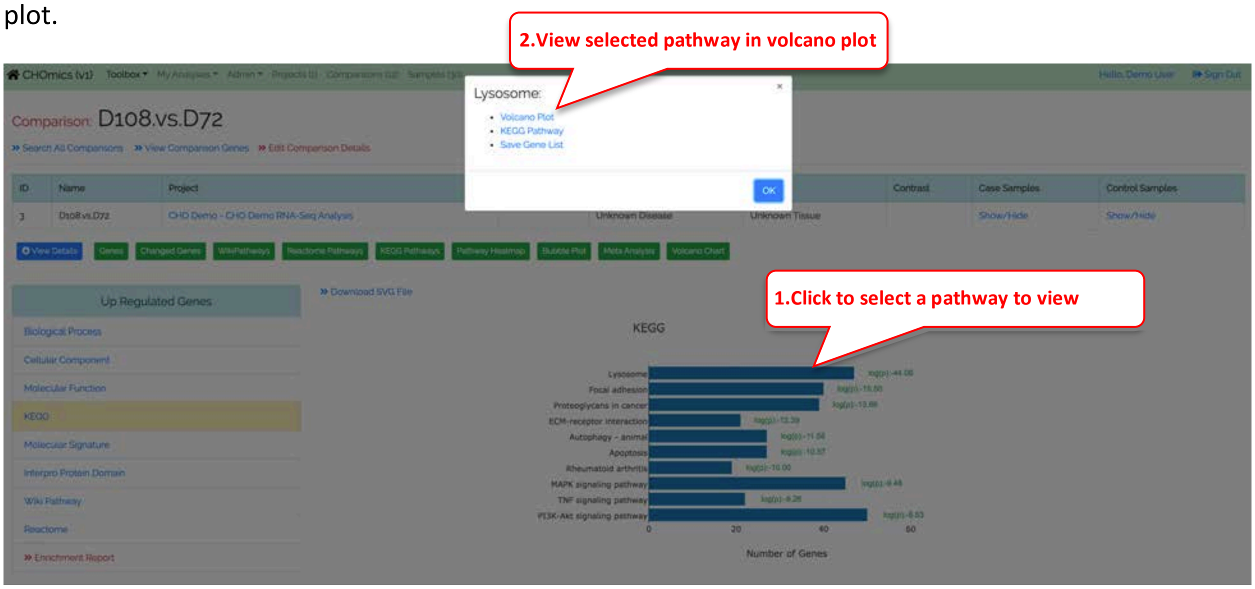
Showing 1 to 10 of 1,001 entries

Previous 1 2 3 4 5 … 101 Next

In the enrichment report, the full list of functional terms are shown by order of p-value.

4.3.2 View Changed Genes from a Functional Term in Volcano Plot

From the bar chat, click a functional term, and you have the option to view these genes in a volcano plot.



2.View selected pathway in volcano plot

1.Click to select a pathway to view

Comparison: D108.vs.D72

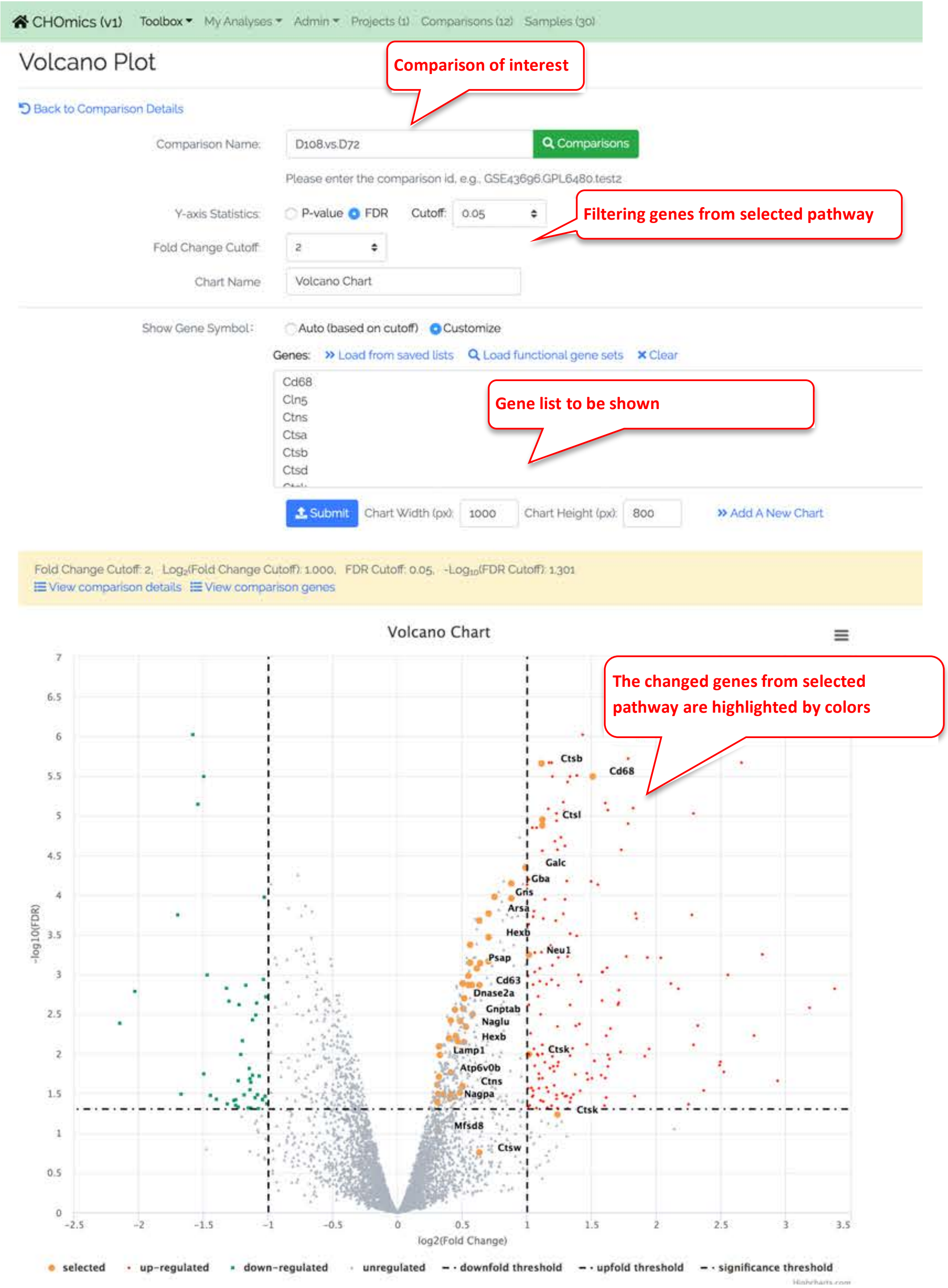
» Search All Comparisons » View Comparison Genes » Edit Comparison Details

ID	Name	Project				Contrast	Case Samples	Control Samples
3	D108.vs.D72	CHO Demo - CHO Demo RNA-Seq Analysis		Unknown Disease	Unknown Tissue		Show/Hide	Show/Hide

Lysosome:

- Volcano Plot
- KEGG Pathway
- Save Gene List

Once you click the link in the popup window, volcano plot will be generated for the comparison with the changed genes from the selected term highlighted.



Volcano Plot

Comparison of interest

Comparison Name: D108.vs.D72

Please enter the comparison id, e.g., GSE43696.GPL6480.test2

Y-axis Statistics: ○ P-value ● FDR Cutoff: 0.05

Filtering genes from selected pathway

Fold Change Cutoff: 2

Chart Name: Volcano Chart

Show Gene Symbol: ○ Auto (based on cutoff) ● Customize

Genes:

Cd68
Cln5
Ctns
Ctsa
Ctsb
Ctsd

Gene list to be shown

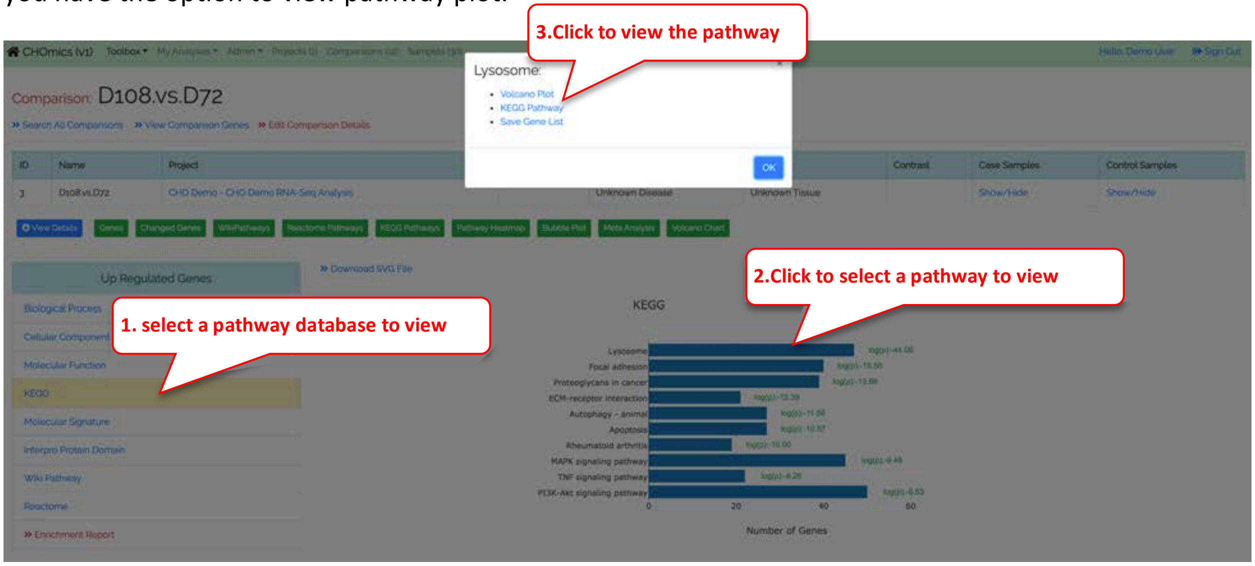
Chart Width (px): 1000 Chart Height (px): 800

Fold Change Cutoff: 2, Log_2(Fold Change Cutoff): 1.000, FDR Cutoff: 0.05, $-Log_{10}$(FDR Cutoff): 1.301

The changed genes from selected pathway are highlighted by colors

4.3.3 View Enriched Pathways Directly from Comparison Details

From the bar chat, if you are viewing KEGG or wikipathway database, clicking the pathway name and you have the option to view pathway plot.



3.Click to view the pathway

2.Click to select a pathway to view

1. select a pathway database to view

Comparison: D108.vs.D72

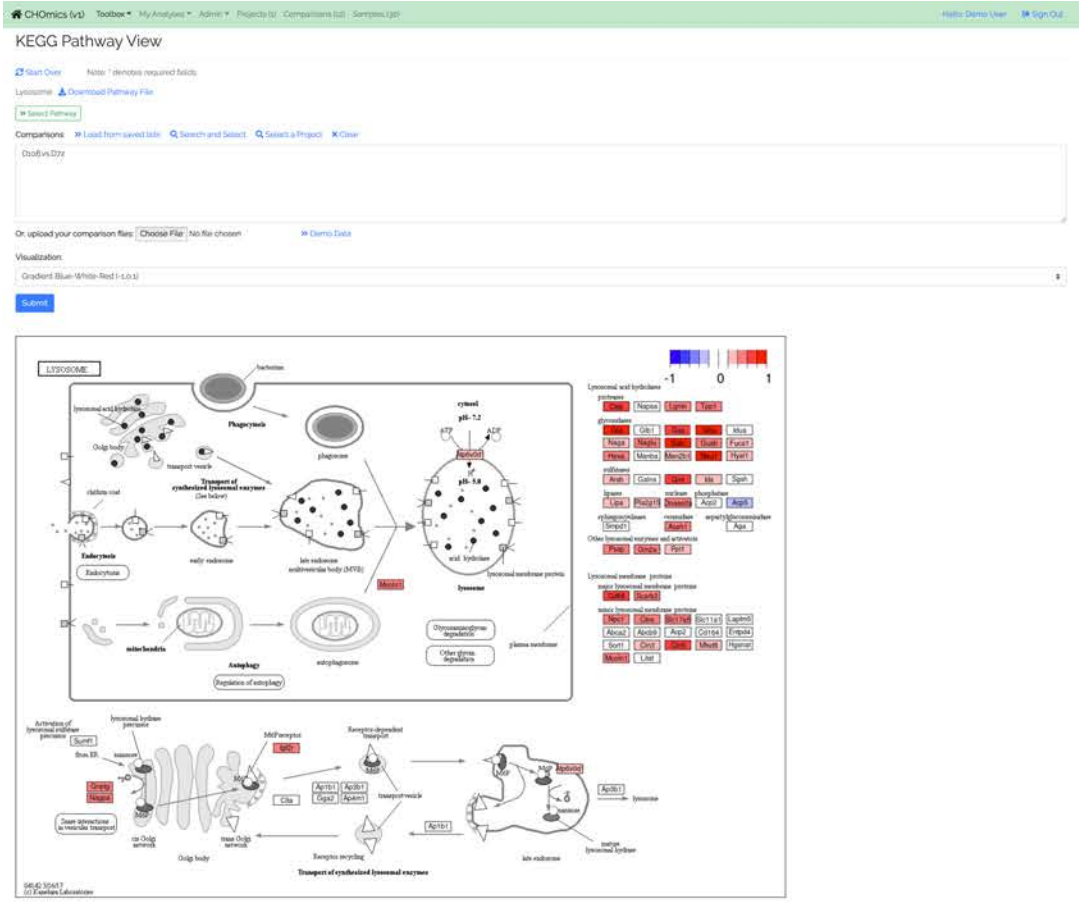
» Search All Comparisons » View Comparison Genes » Edit Comparison Details

ID	Name	Project			Contrast	Case Samples	Control Samples
3	D108.vs.D72	CHO Demo - CHO Demo RNA-Seq Analysis	Unknown Disease	Unknown Tissue		Show/Hide	Show/Hide

Lysosome:

- Volcano Plot
- KEGG Pathway
- Save Gene List

This will automatically open the pathway visualization page, and preload the pathway and comparison. Click submit to view the pathway.



KEGG Pathway View

Note: * denotes required fields

Lysosome Download Pathway File

Comparisons: D108.vs.D72

Or, upload your comparison files: Choose File No file chosen

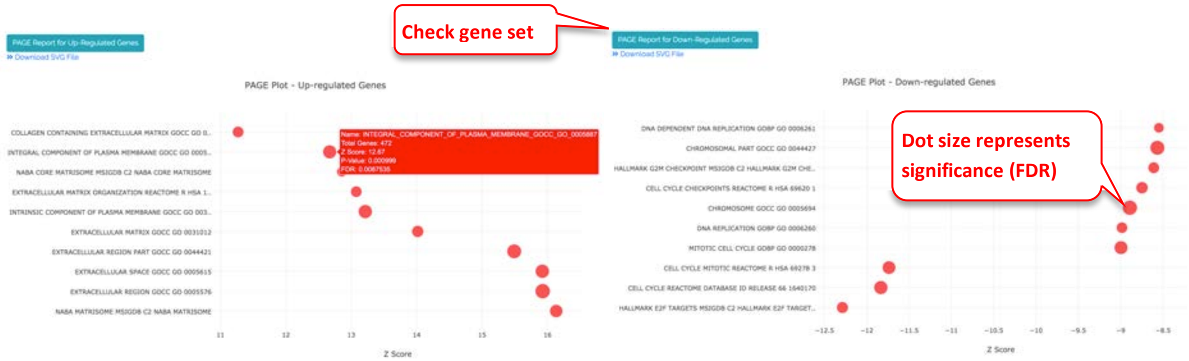
Visualization: Gradient Blue-White-Red (-1,0,1)

Gene Set Enrichment from Ranked Genes

For each comparison, we produce a rank file for all genes using logFC. We use PAGE (Parametric Analysis of Gene Set Enrichment) to identify significant biological changes. PAGE can be more sensitive for comparisons where the logFC is relatively small, but most genes in a functional set show the same direction of change.

The predefined gene sets were from MSigDB.

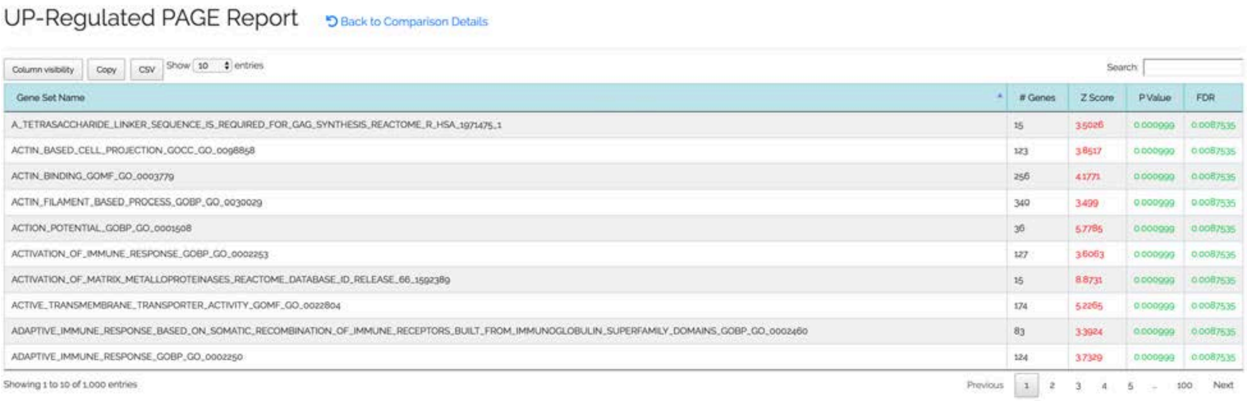
For each comparison, the top up-regulated and down-regulated gene sets are plotted.



Check gene set

Dot size represents significance (FDR)

To view the full list of gene sets, you can click the report for genes as shown in following figure.

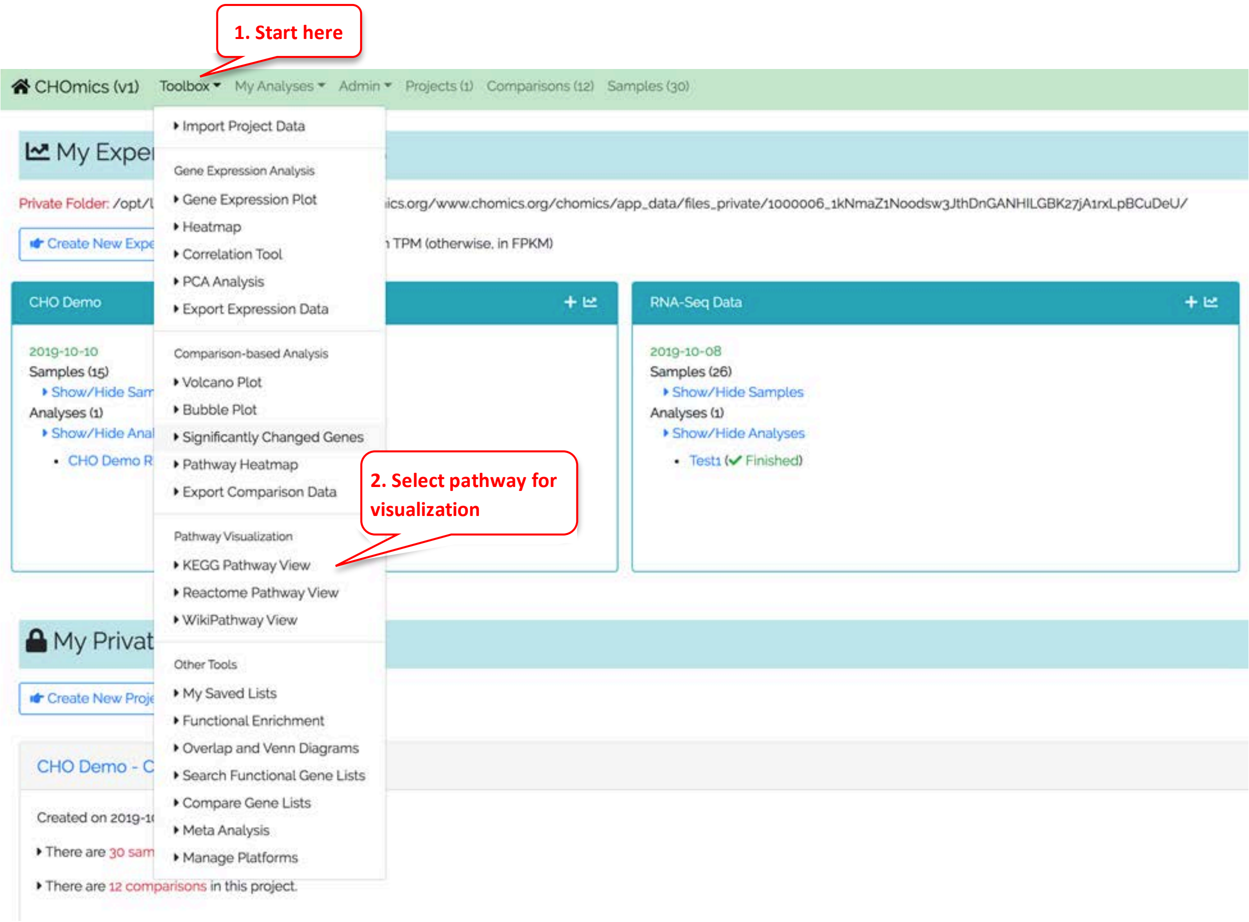


Gene Set Name	# Genes	Z Score	P Value	FDR
A_TETRASACCHARIDE_LINKER_SEQUENCE_IS_REQUIRED_FOR_GAG_SYNTHESIS_REACTOME_R_HSA_1971475_1	15	3.5026	0.000999	0.0087535
ACTIN_BASED_CELL_PROJECTION_GOCC_GO_0098858	123	3.8517	0.000999	0.0087535
ACTIN_BINDING_GOMF_GO_0003779	256	4.1771	0.000999	0.0087535
ACTIN_FILAMENT_BASED_PROCESS_GOBP_GO_0030029	340	3.499	0.000999	0.0087535
ACTION_POTENTIAL_GOBP_GO_0001508	36	5.7785	0.000999	0.0087535
ACTIVATION_OF_IMMUNE_RESPONSE_GOBP_GO_0002253	127	3.6063	0.000999	0.0087535
ACTIVATION_OF_MATRIX_METALLOPROTEINASES_REACTOME_DATABASE_ID_RELEASE_66_1592389	15	8.8731	0.000999	0.0087535
ACTIVE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY_GOMF_GO_0022804	174	5.2265	0.000999	0.0087535
ADAPTIVE_IMMUNE_RESPONSE_BASED_ON_SOMATIC_RECOMBINATION_OF_IMMUNE_RECEPTORS_BUILT_FROM_IMMUNOGLOBULIN_SUPERFAMILY_DOMAINS_GOBP_GO_0002460	83	3.3924	0.000999	0.0087535
ADAPTIVE_IMMUNE_RESPONSE_GOBP_GO_0002250	124	3.7329	0.000999	0.0087535

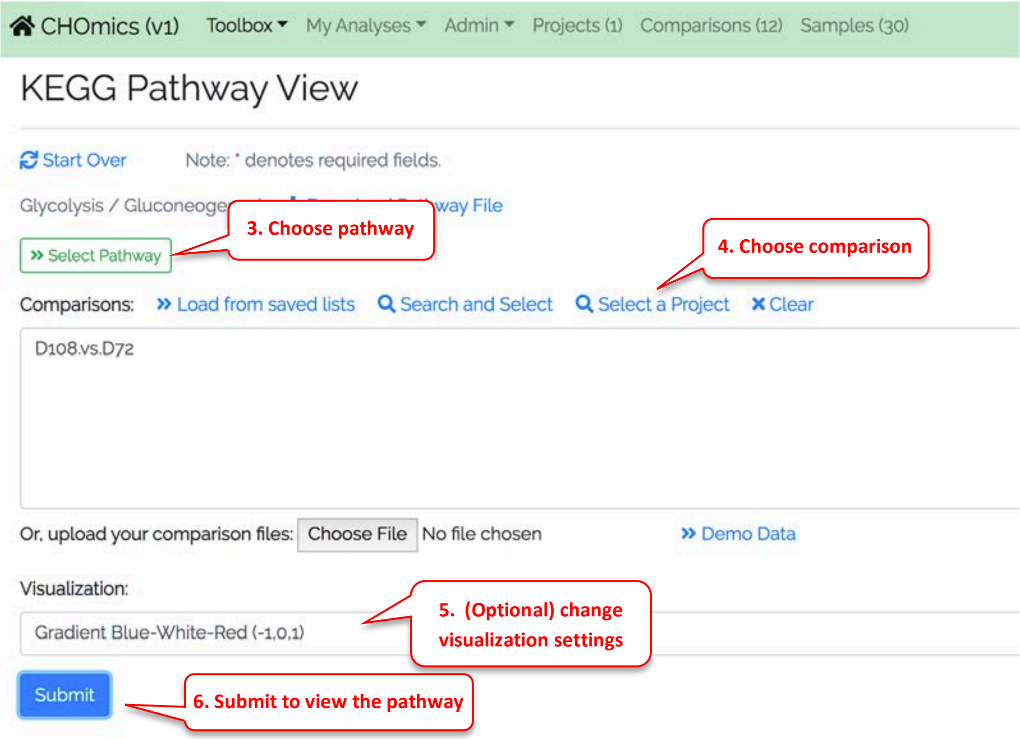
Showing 1 to 10 of 1,000 entries

4.3.4 Multi-layer visualization

If you are interested in a particular pathway, sometimes it is useful to map the RNA-Seq or microarray data to the pathway for visualization.


1. Start here

2. Select pathway for visualization


3. Choose pathway

4. Choose comparison

5. (Optional) change visualization settings

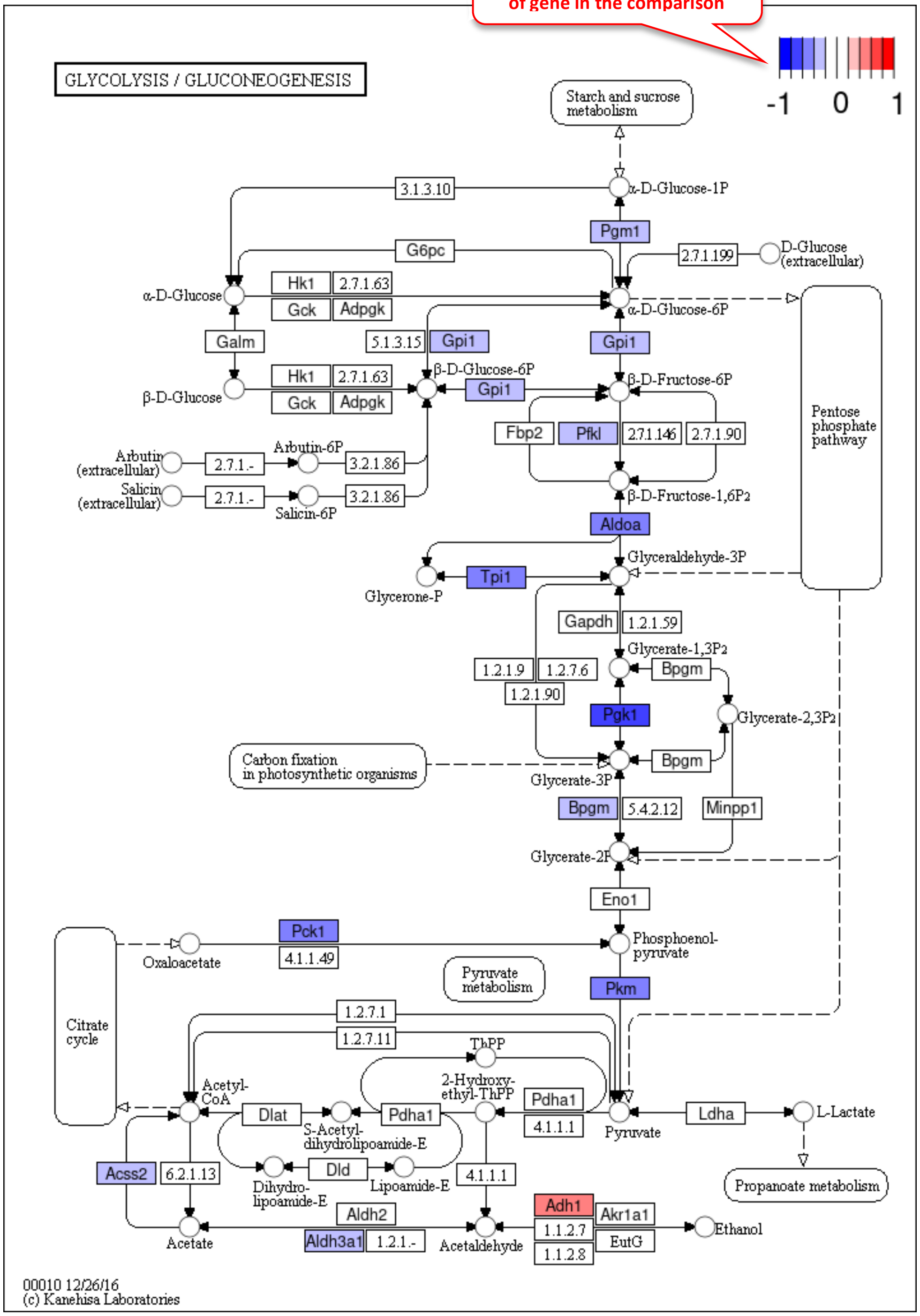
6. Submit to view the pathway

Comparisons: D108.vs.D72

Or, upload your comparison files: No file chosen

Visualization: Gradient Blue-White-Red (-1,0,1)

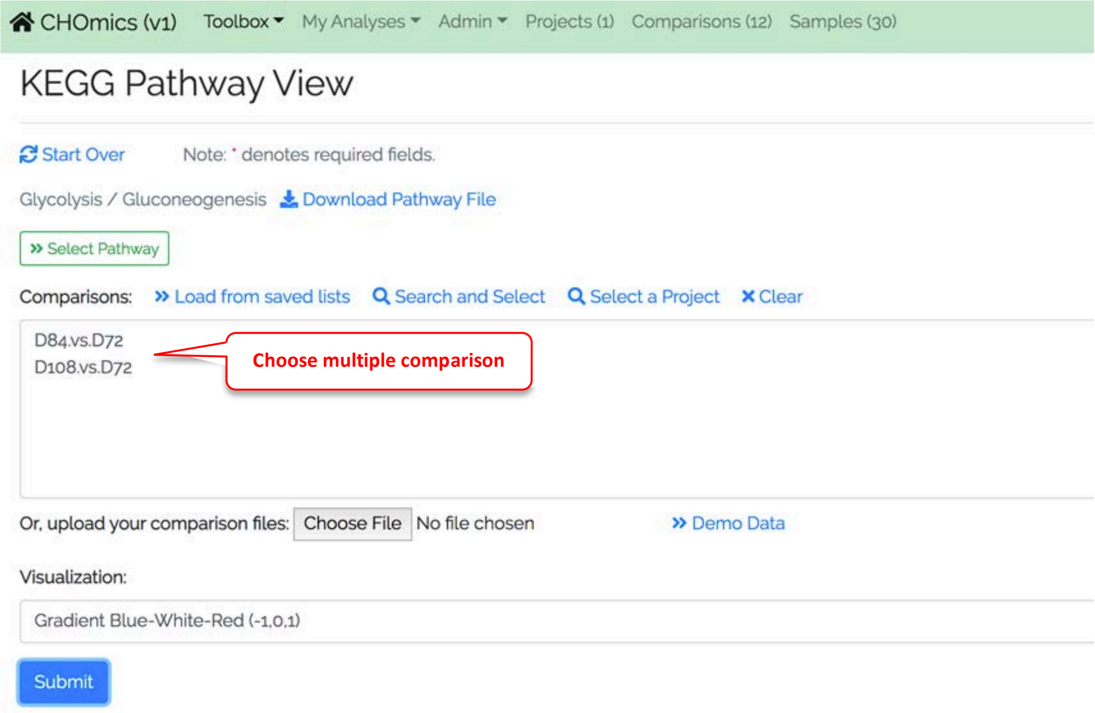
In the pathway plot, typically we use red-blue color scale to show the log2 Fold Change. Blue is down-regulated, red is up-regulated.



Color represents the significance of gene in the comparison

Pathway Plot from Several Comparisons

The user can add multiple comparisons from the pathway plot tool by clicking Add Comparison link. Besides showing log2 Fold Change, the user can also show statistical significance by clicking Enable Second Visualization Columns.



KEGG Pathway View

Note: * denotes required fields.

Glycolysis / Gluconeogenesis

Comparisons:

D84.vs.D72
D108.vs.D72

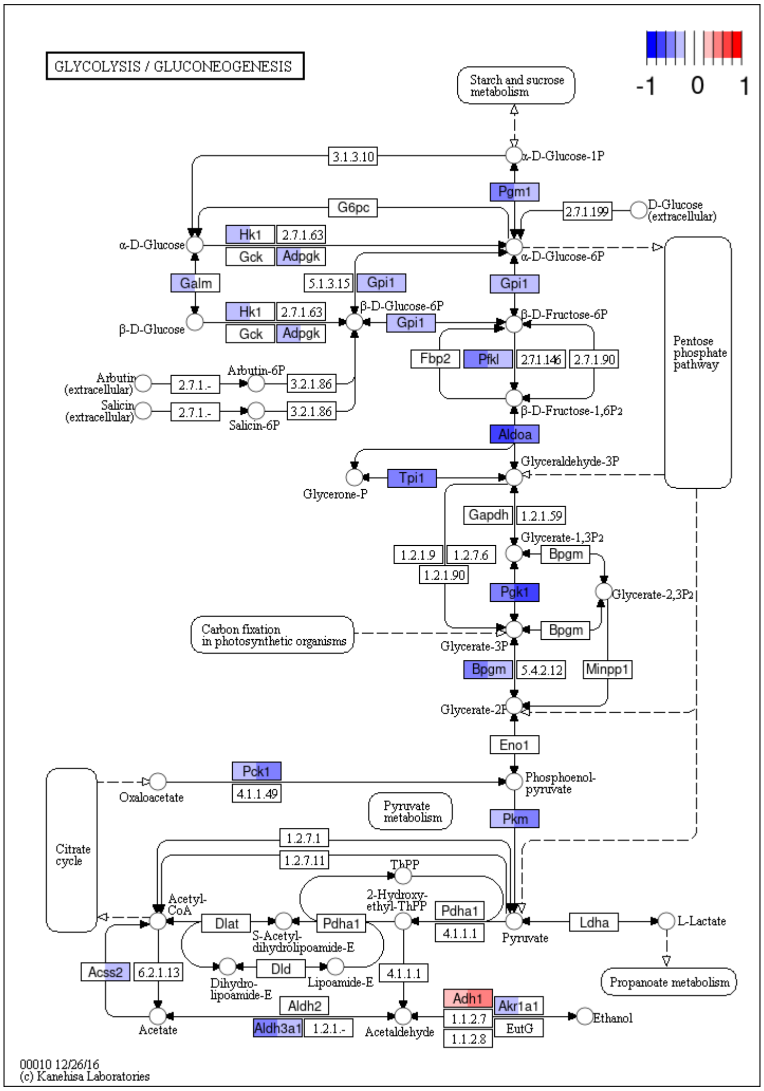
Choose multiple comparison

Or, upload your comparison files:

No file chosen

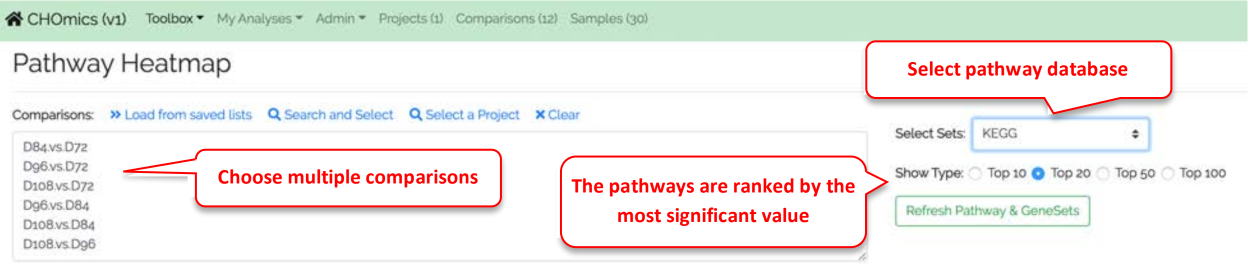
Visualization: Gradient Blue-White-Red (-1,0,1)

The pathway plot will now have multiple color bars corresponding to the different comparisons.



4.3.5 Pathway Heatmap From Comparisons

Users can display the enriched pathways from several related comparisons, and visualize the top enriched pathways across comparisons. Users can mix public data and inhouse comparisons.



Pathway Heatmap

Comparisons: » Load from saved lists Search and Select Select a Project × Clear

D84.vs.D72
Dg6.vs.D72
D108.vs.D72
Dg6.vs.D84
D108.vs.D84
D108.vs.Dg6

Select Sets: KEGG

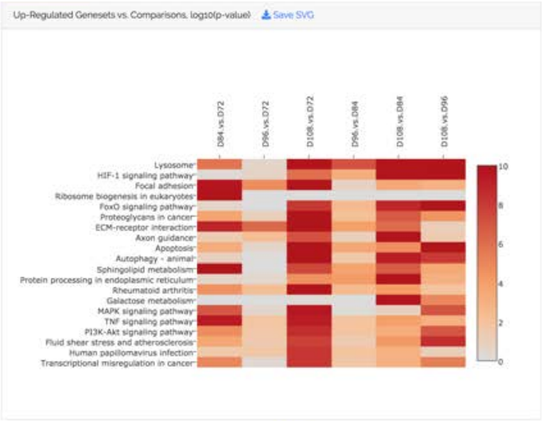
Show Type: ○ Top 10 ● Top 20 ○ Top 50 ○ Top 100

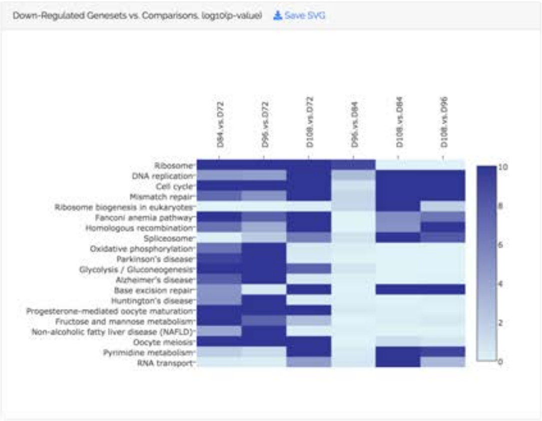
Select pathway database

Choose multiple comparisons

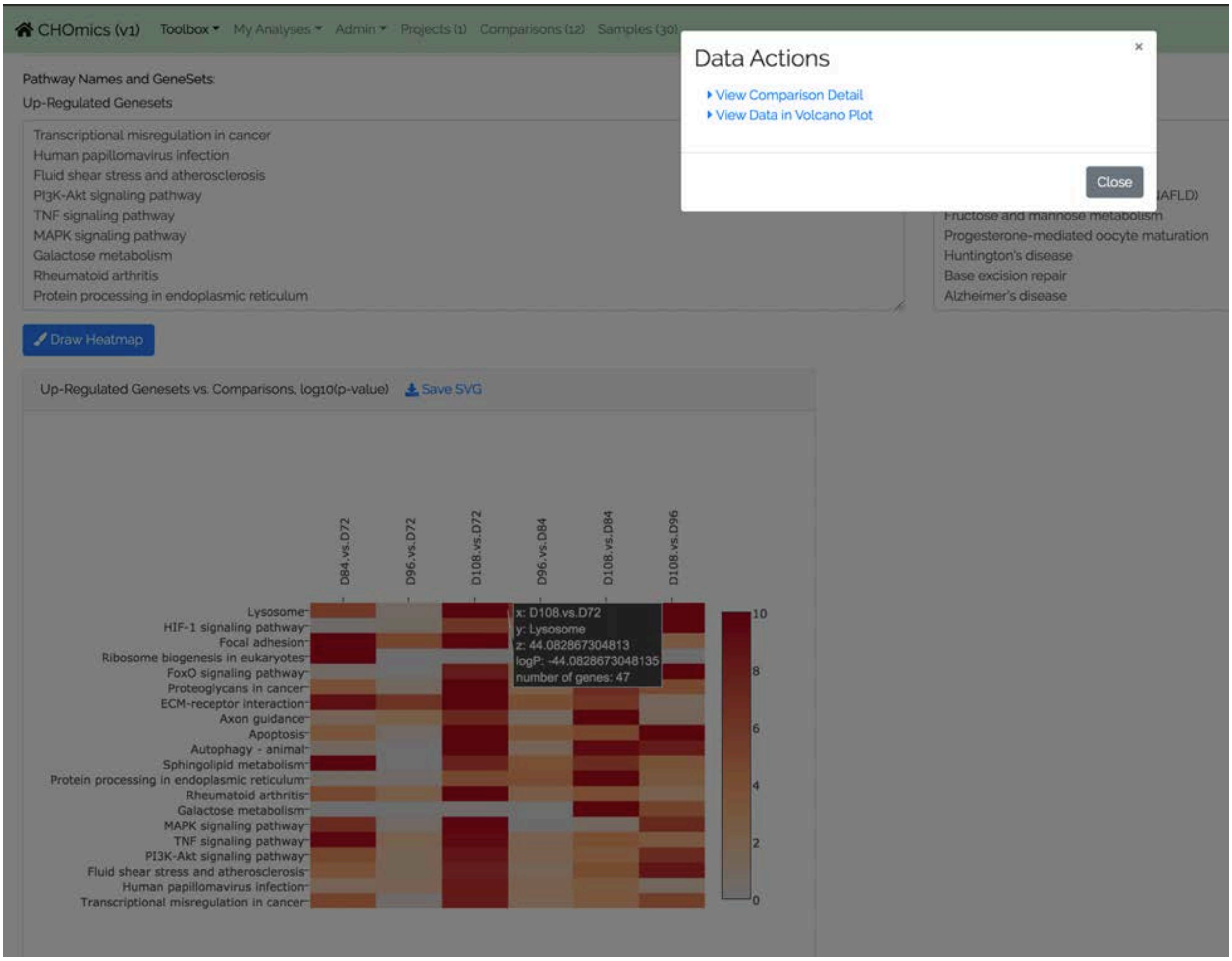
The pathways are ranked by the most significant value

The heatmap shows pathways in rows, comparisons in columns. The statistical significance is color-coded (log P-value, or Z-score). Pathways are sorted by the negative logP values from the highest to the lowest.





From the pathway heatmap, users can click any data point to view details.



Data Actions

- View Comparison Detail
- View Data in Volcano Plot

Pathway Names and GeneSets:

Up-Regulated Genesets

Transcriptional misregulation in cancer
Human papillomavirus infection
Fluid shear stress and atherosclerosis
PI3K-Akt signaling pathway
TNF signaling pathway
MAPK signaling pathway
Galactose metabolism
Rheumatoid arthritis
Protein processing in endoplasmic reticulum

AFLD)
Fructose and mannose metabolism
Progesterone-mediated oocyte maturation
Huntington's disease
Base excision repair
Alzheimer's disease

x: D108.vs.D72
y: Lysosome
z: 44.082867304813
logP: -44.0828673048135
number of genes: 47

5 Customized analysis pipeline

5.1 Use alternative tool or algorithm

The analysis pipeline is modular, each step can be modified by uses to use an alternative method if desired. The users should be familiar with the Linux bash to run the analysis steps and be familiar with php programming to make modification to the source code.

The full analysis pipeline has four steps, and each step is listed in a bash file in the analysis folder in the system.

- step_0.sh FASTQC of raw data
- step_1.sh Alignment to genome
- step_2.sh Gene count
- step_3.sh DEG detection and functional enrichment

These bash files are created by PHP programs chomics/app/bxgenomics/bxgenomics_exe_analysis.php, when users launch analysis pipeline online in a web browser via chomics/app/bxgenomics/analysis.php. For example, the current pipeline uses subread to perform alignment. If users want to modify the pipeline to change it to use the STAR program for alignment, they need the following steps:

1) Install STAR program on the server, prepare STAR index for the CHO genome.

2) Check the commands in step_1.sh, and change the commands as needed. In this case, the subread command (subjunc step) needs to be replaced by the equivalent STAR command. Since STAR can sort the bam files, the samtools sort step can be omitted. Finally, the STAR output file is named as SampleIDAligned.sortedByCoord.out.bam, an extra step is needed to rename it to SampleID.sorted.bam, so step2.sh can output gene count files with the correct sample names.

3) Edit PHP program chomics/app/bxgenomics/bxgenomics_exe_analysis.php, find the part that generates step_1.sh (The section is marked as “Step 1. Alignment with Subread”), and then make changes accordingly.

4) Test the updated system to make sure it works as expected.